# STIC Search Report

## EIC 2100

STIC Database Tracking Number: 161637

**TO: Gwen Liang**
**Location: RND 3B11**
**Art Unit : 2162**
**Tuesday, August 09, 2005**

**Case Serial Number: 10/888895**

**From: Geoffrey St. Leger**
**Location: EIC 2100**
**Randolph-4B31**
**Phone: 23450**

**geoffrey.stleger@uspto.gov**

## Search Notes

Dear Examiner Liang,

Attached please find the results of your search request for application 10/888895. I searched Dialog's patent files, technical databases and general files.

Please let me know if you have any questions.

Regards,

Geoffrey St. Leger

Geoffrey St. Leger
4B31/x23540

from that statistical **information** , a second calculation means 11 that obtains **category** **information** of a **document** in which each **keyword** **appears** by referring to the **document** database and calculates the weight of each **keyword** from the **category** **information** , taking into account the **frequency** of **appearance** of each **keyword** , and a generation means 12 that generates the weight of each keyword by determining a compared value of the degree of importance between the weight the first calculation means 10 calculates and the one the second calculation means 11 calculates, and synthesizing their weight in accordance with the compared value of the degree of importance.

**13/5/4** **(Item 4 from file: 347)**
DIALOG(R)File 347:JAPIO

06144257 **Image available**
AUTOMATIC DOCUMENT CLASSIFICATION DEVICE, LEARNING DEVICE, CLASSIFICATION DEVICE, AUTOMATIC DOCUMENT CLASSIFICATION METHOD, LEARNING METHOD, CLASSIFICATION METHOD AND STORAGE MEDIUM

| | |
|---|---|
| PUB. NO.: | 11-085797 [JP 11085797 A] |
| PUBLISHED: | March 30, 1999 ( **19990330**) |
| INVENTOR(s): | OTANI NORIKO<br>ITO SHIRO<br>SHIBATA SHOGO<br>UEDA TAKANARI<br>IKEDA YUJI |
| APPLICANT(s): | CANON INC |
| APPL. NO.: | 09-250126 [JP 97250126] |
| FILED: | September 01, 1997 (19970901) |
| INTL CLASS: | **G06F-017/30** |

ABSTRACT

PROBLEM TO BE SOLVED: To provide an automatic document classification device which can form a vector space where **topics** are precisely reflected and which can appropriately execute classification.

SOLUTION: The automatic document classification device selects a valid word from a learning document (valid word selection part 103). The number of the valid words contained in respective paragraphs is obtained by referring to the learning **document** and the valid **word** (intra-paragraph valid **word** number calculation part 105). The intra-paragraph cooccurrence **frequency** of the **group** of the respective valid **words** is obtained by using the number of intra-paragraph valid words (intra-paragraph cooccurrence calculation part 107). The valid word vectors of the respective valid words are obtained from obtained intra- paragraph cooccurrence frequency, and the document vectors are obtained on the learning document and the document being a classification object by referring to the valid word vectors. The folder vectors of the respective **categories** , which are obtained from the document vector of the learning document, are compared with the document vector of the document being the classification object. The **category** to which the document being the classification object belongs is decided in accordance with the compared result.

**13/5/9** **(Item 9 from file: 347)**
DIALOG(R)File 347:JAPIO

05211465 **Image available**
METHOD FOR AUTOMATICALLY CLASSIFYING JAPANESE TEXT

PUB. NO.: 08-166965 [JP 8166965 A]
PUBLISHED: June 25, 1996 ( **19960625)**
INVENTOR(s): SUNABA RINTAROU
APPLICANT(s): NIPPON TELEGR & TELEPH CORP <NTT> [000422] (A Japanese
Company or Corporation), JP (Japan)
APPL. NO.: 06-310875 [JP 94310875]
FILED: December 14, 1994 (19941214)
INTL CLASS: [6] **G06F-017/30** ; **G06F-017/27**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)

ABSTRACT

PURPOSE: To automatically classify a newly inputted Japanese text by learning appearance **frequency information** of a **word** (a noun, a verb, an adjective and an adverb) being intrinsic to a **category** and of language expression being equal to a modifier and a word to be modified in a text database which is previously classified into several **categories** .

CONSTITUTION: An automatic classification rule learning part 17 accesses to a learning text storing device 6 and executes learning from the classified text so that anti- **category** language expression importance degree tables 7 and 8 are generated. Then, an automatic text classifying part 18 accesses to the anti- **category** language expression importance degree table 8 as against the text inputted from a user text input device 19 and a classified result is outputted from a classification result display device 20.

**13/5/19 (Item 19 from file: 347)**
DIALOG(R)File 347:JAPIO

03462366 **Image available**
ELECTRONIC DOCUMENT FILING SYSTEM

PUB. NO.: 03-125266 [JP 3125266 A]
PUBLISHED: May 28, 1991 ( **19910528)**
INVENTOR(s): HARAGUCHI SATOSHI
APPLICANT(s): MITSUBISHI ELECTRIC CORP [000601] (A Japanese Company or
Corporation), JP (Japan)
APPL. NO.: 01-264177 [JP 89264177]
FILED: October 11, 1989 (19891011)
INTL CLASS: [5] **G06F-015/40**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)
JOURNAL: Section: P, Section No. 1243, Vol. 15, No. 338, Pg. 41,
August 27, 1991 (19910827)

ABSTRACT

PURPOSE: To quickly and easily turn a document into an electronic form by adding a key word extracting function to a system to extract the key word of a **subject** electronic document file.

CONSTITUTION: A document filling system is provided with a key word dictionary memory 9 which stores a key word dictionary. The contents of the memory 9 are compared with a **subject** file A (B), and the coincidence frequency is counted between the **contents** of the dictionary 9 and the key **words** of a key **word** candidate **group** included in the file A (B). Then the key **word** candidate having high coincidence **frequency** is defined as a key word. Thus a key word is defined to characterize a sentence which extracts a key word out of a document and the using frequency of the key word. Then the subsequent documents can be easily retrieved.

**13/5/20 (Item 20 from file: 347)**

DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

03130473 **Image available**
AUTOMATIC CLASSIFYING DEVICE FOR DOCUMENT

```
PUB. NO.:      02-105973  [JP 2105973  A]
PUBLISHED:     April 18, 1990 ( 19900418)
INVENTOR(s):   KAWAI ATSUO
               NAGATA MASAAKI
               KIMOTO HARUO
APPLICANT(s):  NIPPON TELEGR & TELEPH CORP <NTT> [000422] (A Japanese
               Company or Corporation), JP (Japan)
APPL. NO.:     63-258748  [JP 88258748]
FILED:         October 14, 1988 (19881014)
INTL CLASS:    [5]  G06F-015/40
JAPIO CLASS:   45.4 (INFORMATION PROCESSING -- Computer Applications)
JOURNAL:       Section: P, Section No. 1075, Vol. 14, No. 325, Pg. 18, July
               12, 1990 (19900712)
```

ABSTRACT

PURPOSE: To identify a word in the same set even when a word which has the same concept with a word (field identification word) expressing features by fields, but is different as a character string appears in an unclassified document by using meaning **categories** as a clue to classification.

CONSTITUTION: The meaning **categories** of **words** are noticed and a meaning **category** which **appears** one-sidedly by the fields is used as a new clue to classify **documents** . Namely, the features (field-by-field deviation in **appearance** **frequency** of a **keyword** and the meaning **category** ) are recorded in a field identification **word** point table 3a nd a field identification means **category** point table 4. Consequently, when the word (deviation in expression and homonym) which has the same concept with the field identification **word** representing the features by the fields, but is different as the character string **appears** in the unclassified **document** , the meaning **category** is used to identify the **word** in the same set, thereby obtaining the clue to the classification.

**13/5/23 (Item 2 from file: 350)**
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

014834097 **Image available**
WPI Acc No: 2002-654803/200270
XRPX Acc No: N02-517336

**Interactive classification and analysis method for textual data in helpdesk service, involves displaying table including name, cohesion score and distinctness score for each cluster of documents**

Patent Assignee: INT BUSINESS MACHINES CORP (IBMC )
Inventor: KREULEN J T; MODHA D S; SPANGLER W S; STRONG H R
Number of Countries: 001 Number of Patents: 001
Patent Family:

```
Patent No     Kind   Date     Applicat No     Kind   Date      Week
US 6424971     B1  20020723  US 99429650      A    19991029  200270  B
```

Priority Applications (No Type Date): US 99429650 A 19991029
Patent Details:

```
Patent No  Kind Lan Pg   Main IPC     Filing Notes
US 6424971    B1    15 G06F-017/30
```

Abstract (Basic): US 6424971 B1

NOVELTY - The dictionary including a subset of **words** contained in a **document** set and count of frequency **occurrence** of each **word** in the **document** set are generated. The set of **documents** are partitioned into multiple **clusters** for which the name and centroid in the dictionary space are generated. The cohesion and distinctness scores are generated for each cluster. A table including the name, cohesion score and distinctness score for each cluster is displayed.

DETAILED DESCRIPTION - INDEPENDENT CLAIMS are included for the following:

(1) Interactive classification and analysis system; and

(2) Computer program product for interactive classification and analysis.

USE - For interactive classifying and analyzing textual data in helpdesk service.

ADVANTAGE - Clustering of documents enables a user to determine the content of documents in the cluster without having to look at all of the documents. This saves the user's considerable time and ultimately reduces expenses. Enables identifying candidate helpdesk problem **categories** that are most amendable to automated solutions and hence improves the efficiency of the helpdesk operation.

DESCRIPTION OF DRAWING(S) - The figure shows a flow diagram of the interactive classification and analysis process.

pp; 15 DwgNo 4/8

Title Terms: INTERACT; CLASSIFY; ANALYSE; METHOD; TEXT; DATA; SERVICE; DISPLAY; TABLE; NAME; COHERE; SCORE; SCORE; CLUSTER; DOCUMENT
Derwent Class: T01
International Patent Class (Main): **G06F-017/30**
File Segment: EPI

**13/5/24 (Item 3 from file: 350)**
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

013617938 **Image available**
WPI Acc No: 2001-102146/200111
Related WPI Acc No: 2003-265503; 2003-379236
XRPX Acc No: N01-075883

**On-line query supporting method for e-com in Internet, involves mapping terms in super** category **to documents** category **and weighting terms in received query to rank and select relevant super** category **term from list**

Patent Assignee: GTE LAB INC (SYLV ); VERIZON LAB INC (VERI-N)
Inventor: PONTE J
Number of Countries: 092 Number of Patents: 003
Patent Family:

| Patent No | Kind | Date | Applicat No | Kind | Date | Week | |
|---|---|---|---|---|---|---|---|
| WO 200058863 | A1 | 20001005 | WO 2000US8450 | A | 20000330 | 200111 | B |
| AU 200043280 | A | 20001016 | AU 200043280 | A | 20000330 | 200111 | |
| US 6826559 | B1 | 20041130 | US 99283268 | A | 19990331 | 200479 | |

Priority Applications (No Type Date): US 99283268 A 19990331; US 99282730 A 19990331
Patent Details:

| Patent No | Kind | Lan | Pg | Main IPC | Filing Notes |
|---|---|---|---|---|---|
| WO 200058863 | A1 | E | 186 | G06F-017/10 | |

Designated States (National): AE AG AL AM AT AU AZ BA BB BG BR BY CA CH CN CR CU CZ DE DK DM DZ EE ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MA MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT TZ UA UG UZ VN YU ZA ZW
Designated States (Regional): AT BE CH CY DE DK EA ES FI FR GB GH GM GR IE IT KE LS LU MC MW NL OA PT SD SE SL SZ TZ UG ZW

| Patent No | Kind | Lan | Pg | Main IPC | Filing Notes |
|---|---|---|---|---|---|
| AU 200043280 | A | | | G06F-017/10 | Based on patent WO 200058863 |
| US 6826559 | B1 | | | G06F-017/30 | |

Abstract (Basic): WO 200058863 A1

NOVELTY - A list of super **category** terms that are linked to specific application is prepared based on the **category** of documents to be searched and the listed terms are mapped against document **category** . The **category** is retrieved based on terms in user input query. The query is then modified and terms in the query are weighted to determine most relevant super **category** term by ranking method.

DETAILED DESCRIPTION - The weighting of the modified query is performed by computing sum of **term** frequency and inverse **document** frequency of each **term** in the super **category** **terms** list. The inverse **document** **frequency** is set as high value, when **terms** **appearing** in the **category** is manually mapped against the super **category** , when compared to the **terms** that are automatically mapped. INDEPENDENT CLAIMS are also included for the following:

(a) computer program for ranking super **categories** used for data query;

(b) program for searching document;

(c) program for establishing super **category** terms list

USE - For displaying on-line banner advertisements for user query for e-com in Internet.

ADVANTAGE - The user's query can be cached and subset or superset of cached data can be referred for subsequent queries which enhances the response for subsequent user queries.

DESCRIPTION OF DRAWING(S) - The figure shows the block diagram of software links of on-line query tool.

pp; 186 DwgNo 4/71

Title Terms: LINE; QUERY; SUPPORT; METHOD; MAP; TERM; SUPER; **CATEGORY** ; DOCUMENT; **CATEGORY** ; WEIGHT; TERM; RECEIVE; QUERY; RANK; SELECT; RELEVANT; SUPER; **CATEGORY** ; TERM; LIST

Derwent Class: T01; W01

International Patent Class (Main): **G06F-017/10** ; **G06F-017/30**

International Patent Class (Additional): **G06F-005/14** ; G06K-009/72; H04N-007/14

File Segment: EPI

**13/5/25 (Item 4 from file: 350)**

DIALOG(R)File 350:Derwent WPIX

(c) 2005 Thomson Derwent. All rts. reserv.

012552401 **Image available**

WPI Acc No: 1999-358507/ **199931**

XRPX Acc No: N01-054163

Topic **words establishing method in computer, involves selecting words** belonging to different pre-established keyword classes from document **keywords as** topic **words**

Patent Assignee: UNIV HONG KONG CHINESE LANGUAGE (UYHK-N); UNIV HONG KONG CHINESE LANGUAGE (UYHO-N); UNIV CHINESE HONG KONG (UYCH-N)

Inventor: QIN A; WONG W S

Number of Countries: 002 Number of Patents: 003

Patent Family:

| Patent No | Kind | Date | Applicat No | Kind | Date | Week | |
|---|---|---|---|---|---|---|---|
| CN 1211769 | A | 19990324 | CN 98102672 | A | 19980626 | 199931 | B |
| US 6128613 | A | 20001003 | US 9750818 | P | 19970626 | 200109 | |
| | | | US 9869618 | A | 19980429 | | |
| CN 1096038 | C | 20021211 | CN 98102672 | A | 19980626 | 200528 | |

Priority Applications (No Type Date): US 9869618 A 19980429; US 9750818 P 19970626

Patent Details:

| Patent No | Kind | Lan | Pg | Main IPC | Filing Notes |
|---|---|---|---|---|---|
| CN 1211769 | A | | 1 | G06F-017/30 | |
| US 6128613 | A | | 16 | G06F-017/30 | Provisional application US 9750818 |
| CN 1096038 | C | | | G06F-017/30 | |

Abstract (Basic): US 6128613 A

NOVELTY - A portion of document including words is accepted from a data input device to determine several document keywords. Each of the keywords are classified into one of pre-established keyword classes, words belonging to a different pre-established keyword classes are selected from the document keywords as **topic** words.

DETAILED DESCRIPTION - The **words** belonging to different pre-established **keyword** classes are selected to minimize entropy-based cost function on proposed **topic** **words** . The cost function is a metric of dissimilarity between statistical distribution of likelihood of **appearance** by several **document** **keywords** in a typical **document** . An INDEPENDENT CLAIM is also included for **topic** **words** establishing system.

USE - For indexing and retrieval of information from computer databases for recognition of character-based language script and letter based romanized language script, including Chinese, Korean and Japanese as character based language examples and English, French, Spanish, German and Russian as romanized language examples.

ADVANTAGE - Enables reduced storage for index structures and improves recall and precision, even when applied to large databases. Enables document indexing and ordered retrieval of establishing **topic** words to represent each document.

DESCRIPTION OF DRAWING(S) - The figure shows the system for generating statistical characteristics of database.

pp; 16 DwgNo 2/6

Title Terms: **TOPIC** ; WORD; ESTABLISH; METHOD; COMPUTER; SELECT; WORD; BELONG; PRE; ESTABLISH; KEYWORD; CLASS; DOCUMENT; KEYWORD; **TOPIC** ; WORD
Derwent Class: T01
International Patent Class (Main): **G06F-017/30**
File Segment: EPI

**13/5/26 (Item 5 from file: 350)**
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

012337083 **Image available**
WPI Acc No: 1999-143190/ **199912**
XRPX Acc No: N99-104009

**Information mining tool for processing documents stored in database to extract** topics **related to documents - determines determines** topic **trend parameter and parameter corresponding to number of** documents **in which** topic appears **, then determines number of** appearances **in text of** words **corresponding to given** topic

Patent Assignee: DATOPS SA (DATO-N)
Inventor: GAY L; MASSIOT O
Number of Countries: 021 Number of Patents: 001
Patent Family:

```
Patent No     Kind   Date      Applicat No    Kind   Date      Week
WO 9905614     A1   19990204  WO 98IB1123     A    19980723  199912  B
```

Priority Applications (No Type Date): US 9753546 P 19970723
Patent Details:

```
Patent No    Kind Lan Pg   Main IPC     Filing Notes
WO 9905614    A1 E   30 G06F-017/30
   Designated States (National): IL JP US
   Designated States (Regional): AT BE CH CY DE DK ES FI FR GB GR IE IT LU
   MC NL PT SE
```

Abstract (Basic): WO 9905614 A

NOVELTY - The information mining tool includes a mining mechanism for processing documents (11) stored in a database (14) in order to extract the **topics** to which these documents relate. A further mechanism determines parameters which relate to the evolution with time of the **topics** .

USE - For enhancing the intelligence with which information can be analysed and better delivered to customers.

ADVANTAGE - Provides intelligent searching tool providing quantitive and qualitative analysis on wide range of sources from structured to unstructured information. DESCRIPTION OF DRAWING(S) - The drawing shows a schematic drawing I;;ustrating the architecture of the system. (11) full text documents; (14) index database.

Dwg.2/3

Title Terms: INFORMATION; MINE; TOOL; PROCESS; DOCUMENT; STORAGE; DATABASE; EXTRACT; **TOPIC** ; RELATED; DOCUMENT; DETERMINE; DETERMINE; **TOPIC** ; TREND ; PARAMETER; PARAMETER; CORRESPOND; NUMBER; DOCUMENT; **TOPIC** ; APPEAR; DETERMINE; NUMBER; APPEAR; TEXT; WORD; CORRESPOND; **TOPIC**
Derwent Class: T01; W01
International Patent Class (Main): **G06F-017/30**
File Segment: EPI

**13/5/33 (Item 12 from file: 350)**
DIALOG(R)File 350:Derwent WPIX
(c) 2005 Thomson Derwent. All rts. reserv.

004247735
WPI Acc No: 1985-074613/ **198512**
XRPX Acc No: N85-055811

**Automatically locating in text manuscript** subjects **from list - by determining whether each word of segment of text is included in list**

Patent Assignee: AMERICAN TELEPHONE & TELEGRAPH CO (AMTT )
Inventor: RAYE C
Number of Countries: 012 Number of Patents: 007
Patent Family:

| Patent No | Kind | Date | Applicat No | Kind | Date | Week | |
|---|---|---|---|---|---|---|---|
| WO 8501135 | A | 19850314 | WO 84US1228 | A | 19840806 | 198512 | B |
| EP 155284 | A | 19850925 | EP 84903210 | A | 19840806 | 198539 | |
| JP 60502175 | W | 19851212 | JP 84503183 | A | 19840806 | 198605 | |
| US 4580218 | A | 19860401 | | | | 198616 | |
| EP 155284 | B | 19901122 | | | | 199047 | |
| DE 3483651 | G | 19910103 | | | | 199102 | |
| IT 1205650 | B | 19890323 | | | | 199129 | |

Priority Applications (No Type Date): US 83530387 A 19830908
Cited Patents: 5.Jnl.Ref; US 4358824; EP 75903
Patent Details:

| Patent No | Kind | Lan | Pg | Main IPC | Filing Notes |
|---|---|---|---|---|---|
| WO 8501135 | A | E | 35 | | |

Designated States (National): JP
Designated States (Regional): AT BE CH DE FR GB LU NL SE

| Patent No | Kind | Lan | Pg | Main IPC | Filing Notes |
|---|---|---|---|---|---|
| EP 155284 | A | E | | | |

Designated States (Regional): BE DE FR GB NL SE

| Patent No | Kind | Lan | Pg | Main IPC | Filing Notes |
|---|---|---|---|---|---|
| EP 155284 | B | | | | |

Designated States (Regional): BE DE FR GB NL SE

Abstract (Basic): WO 8501135 A

The method is performed by deciding for each word of a predetermined segment of text whether or not it is contained in any word position in at least one **subject** of the list and determining for each **subject** containing a segment word whether or not all other words of each **subject** are contained in a portion of the manuscript of predetermined size, and including the same segment.

A **subject** , formed in the determining step to have all of its words in a segment, is recorded with an indication of the location of such **subject** and segment in the manuscript. The deciding, determining and recording steps are repeated for at least partially different segments of text.

ADVANTAGE - Obviates need for user to decide which combination of **words** constitute an **occurrence** of a particular **subject** and try to locate every **occurrence** of every **subject** in the **document** .

0/7

Title Terms: AUTOMATIC; LOCATE; TEXT; MANUSCRIPT; **SUBJECT** ; LIST;

**14/5/16 (Item 16 from file: 347)**
DIALOG(R)File 347:JAPIO

05667521 **Image available**
WORD CLASSIFICATION PROCESSING METHOD AND DEVICE THEREFOR, AND VOICE RECOGNIZER

PUB. NO.: 09-282321 [JP 9282321 A]
PUBLISHED: October 31, 1997 ( **19971031**)
INVENTOR(s): SHIODA AKIRA
IIDA HITOSHI
APPLICANT(s): ATR ONSEI HONYAKU TSUSHIN KENKYUSHO KK [000000] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 08-198950 [JP 96198950]
FILED: July 29, 1996 (19960729)
INTL CLASS: [6] **G06F-017/28** ; **G06F-017/27** ; G10L-003/00
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications); 42.5 (ELECTRONICS -- Equipment)
JAPIO KEYWORD:R108 (INFORMATION PROCESSING -- Speech Recognition & Synthesis)

ABSTRACT

PROBLEM TO BE SOLVED: To obtain a word classification result having a well-balanced hierarchical structure by classifying plural words into plural classes in the form of a binary tree having hierachized lower, intermediate and upper layers.

SOLUTION: The word classification processing part 20 classifies the words included in the text **data** stored in a text **data** memory 10 by assigning the **words** of comparatively low **appearance frequency** and the **words** of high rates to be adjacent to the same **word** in the same **classes** respectively. Then, the part 20 classes the word classification result into the intermediate, upper and lower layers. Then, the words are classified in order of intermediate, upper and lower layers and based on the prescribed average mutual information content, i.e., a global (overall) cost function set for all words included in the text data. The classified words are stored in a word dictionary memory 11 in the form of a word dictionary. In such word classification processing, it is possible to obtain the word classification result that has a well-balanced hierarchical structure and also is globally optimized.

**14/5/18 (Item 18 from file: 347)**
DIALOG(R)File 347:JAPIO

05538257 **Image available**
DOCUMENT CLASSIFYING DEVICE AND DOCUMENT GROUP DIVIDING METHOD

PUB. NO.: 09-153057 [JP 9153057 A]
PUBLISHED: June 10, 1997 ( **19970610**)
INVENTOR(s): TANOSAKI YASUO
APPLICANT(s): TOSHIBA CORP [000307] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 07-311056 [JP 95311056]
FILED: November 29, 1995 (19951129)
INTL CLASS: [6] **G06F-017/30** ; **G06F-017/21**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)
JAPIO KEYWORD:R139 (INFORMATION PROCESSING -- Word Processors)

ABSTRACT

PROBLEM TO BE SOLVED: To automatically classify a large amount of document data with efficiency by using a word which is present in common to only a 1st group and a word which is present in common to only a 2nd group as key words for group division.

SOLUTION: This document classifying device is equipped with constituent elements such as an input device 1 for inputting characters and commands, a display device 2 which consists of a display device such as a CRT display and displays a list of groups and document contents, an external storage device which consists of a hard disk device, etc., a control unit 4 which consists of a CPU and a memory and controls the whole device, and a communication device 5. Then the word which is present in common to only the 1st group and the **word** which is present in common to only the 2nd **group** are used as the key **words** for **group** division, and general **words** which **appear** in many **documents** are not used as key **words** for division, thereby enabling automatic and efficient classification.

**14/5/19 (Item 19 from file: 347)**
DIALOG(R)File 347:JAPIO

05505636 **Image available**
WORD COLLATION METHOD

PUB. NO.: 09-120436 [JP 9120436 A]
PUBLISHED: May 06, 1997 ( **19970506)**
INVENTOR(s): MARUKAWA KATSUMI
SHIMA YOSHIHIRO
FUJISAWA HIROMICHI
HANANOI TOSHIHIRO
SHIMOKAWABE HIROAKI
SUGIMOTO TAKEYUKI
KADOTA AKIZO
KAWAGUCHI HISAMITSU
APPLICANT(s): HITACHI LTD [000510] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 08-265739 [JP 96265739]
FILED: October 07, 1996 (19961007)
INTL CLASS: [6] G06K-009/72; **G06F-017/22**
JAPIO CLASS: 45.3 (INFORMATION PROCESSING -- Input Output Units); 45.4 (INFORMATION PROCESSING -- Computer Applications)
JAPIO KEYWORD:R107 (INFORMATION PROCESSING -- OCR & OMR Optical Readers)

ABSTRACT

PROBLEM TO BE SOLVED: To improve the correct reading rate of only KANJI (Chinese character) by initializing only a flag table.

SOLUTION: When a character is recognized, an initialization part 104 of a word collation part 103 is started to initialize a flag table. Then a generation part 105 is started to generate a flag table and a cost table after description of the necessary information. Furthermore, a cost calculation part 106 of a word collation part 103 is started to read the words out of a word dictionary 107 to input them to the flag table based on plural pointer tables. Then the transition is given by an input word and the cost of this word is calculated. A compound word processing part 109 of a postprocessing part 108 processes a compound word and does not process a single candidate word. Then an evaluation part 110 performs the evaluation to rearrange the **words** based on the **information** on the cost of a candidate **word** **group** or a candidate **word** string **group** , the **frequency** added to the **words** , etc.

**14/5/21 (Item 21 from file: 347)**
DIALOG(R)File 347:JAPIO

05308010 **Image available**
AUTOMATIC DOCUMENT CLASSIFICATION SYSTEM

PUB. NO.: 08-263510 [JP 8263510 A]
PUBLISHED: October 11, 1996 ( **19961011)**
INVENTOR(s): RI KO
ABE NAOKI
APPLICANT(s): NEC CORP [000423] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 07-065722 [JP 9565722]
FILED: March 24, 1995 (19950324)
INTL CLASS: [6] **G06F-017/30** ; **G06F-017/27**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)

ABSTRACT

PURPOSE: To perform a precise classifying process so that a classification which is originally not identical is not regarded as identical one by classifying a document unequivocally according to a logically grounded information reference.

CONSTITUTION: This system consists of a statistical processing part 1 and an automatic document classification part 2. The statistical processing part 1 extracts words from plural inputted documents, statistically processes (totalizes) the appearance frequencies of the extracted words in the documents, and generate the appearance frequency vectors of the words, and the automatic **document** classification part 2 classifies the **documents** by using the specific **information** reference by using the **appearance** **frequency** vectors of the **words** in the **document** as **data** . Namely, the **documents** are divided into **clusters** according to the **information** reference by using the **appearance** **frequency** vectors of the **words** in the **documents** to be classified and repeatedly divides them to classify the **documents** unequivocally. Thus, the **documents** can easily be classified on the basis of mathematical statistics and information theory while high precision is maintained.

**14/5/22 (Item 22 from file: 347)**
DIALOG(R)File 347:JAPIO

05307995 **Image available**
AUTOMATIC WORD SORTING SYSTEM

PUB. NO.: 08-263495 [JP 8263495 A]
PUBLISHED: October 11, 1996 ( **19961011)**
INVENTOR(s): RI KO
APPLICANT(s): NEC CORP [000423] (A Japanese Company or Corporation), JP (Japan)
GIJUTSU KENKYU KUMIAI SHINJOHO SHIYORI KAIHATSU KIKO [000000] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 07-065716 [JP 9565716]
FILED: March 24, 1995 (19950324)
INTL CLASS: [6] **G06F-017/28**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications); 30.2 (MISCELLANEOUS GOODS -- Sports & Recreation)

ABSTRACT

PURPOSE: To form a thesaurus for processing a natural language at high speed by sorting **words** by repeating division into **clusters** while using the cooccurrence **frequency** vectors of the **words** of sorting objects corresponding to an **information** quantity reference.

CONSTITUTION: A statistical processing part 1 extracts words from an inputted document, totalizes (sums up) the cooccurrence frequency between the extracted word and the specified context of that word and prepares the cooccurrence frequency vector of the word. On the other hand, an automatic word sorting part 2 sorts the words while using the coccurrence frequency

vector prepared by the statistic processing part 1 and outputs the thesaurus for sorting those words. When sorting the words with the automatic word sorting part 2 in this case, first of all, the word group of the sorting object is divided into two clusters, the relation (full description length) of two clusters at such a time is found, the the words of two clusters are exchanged so that this relation can be minimized corresponding to the prescribed information quantity reference. Then, clustering is performed again to two provided clusters and its division is performed until they can not be divided any more

**14/5/23 (Item 23 from file: 347)**
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

05197621 **Image available**
METHOD AND DEVICE FOR DOCUMENT INFORMATION CLASSIFICATION

PUB. NO.: 08-153121 [JP 8153121 A]
PUBLISHED: June 11, 1996 ( **19960611)**
INVENTOR(s): MORITA TAKAKO
TONO JUNICHI
MATSUDA YOSHIKI
HASHIMOTO TETSUYA
APPLICANT(s): HITACHI LTD [000510] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 07-231033 [JP 95231033]
FILED: September 08, 1995 (19950908)
INTL CLASS: [6] **G06F-017/30 ; G06F-012/00**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications); 45.2 (INFORMATION PROCESSING -- Memory Units)

ABSTRACT

PURPOSE: To provide a method and device for document **information** classification which classify a **document** group without depending upon a prescribed classification system by using a key **word** given to the **document group** or a **word appearing** in a **document** and rearranges classification results hierarchically.

CONSTITUTION: A data management part 101 manages the document group in a document DB 107 and a group of key words, given to respective documents, in a key word DB 108. A document classification part 103 classifies the documents on the basis of the individual key words and stores them in folders. Then, folders having similar document groups are integrated. For the integration, it is judged whether the integration is effective or not. It is judged whether or not further classifications can be made in folders that are left without being integrated, thereby generating a hierarchical classification system. The classification results are outputted on a CRT 109 by a classification output part 104 to provide an environment wherein a user can read the classification results out

**14/5/25 (Item 25 from file: 347)**
DIALOG(R)File 347:JAPIO
(c) 2005 JPO & JAPIO. All rts. reserv.

05027282 **Image available**
JUDGEMENT METHOD FOR KEYWORD

PUB. NO.: 07-319882 [JP 7319882 A]
PUBLISHED: December 08, 1995 ( **19951208)**
INVENTOR(s): ARITA MASATAKE
APPLICANT(s): NEC CORP [000423] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 06-106986 [JP 94106986]

FILED: May 20, 1994 (19940520)
INTL CLASS: [6] **G06F-017/30**
JAPIO CLASS: 45.4 (INFORMATION PROCESSING -- Computer Applications)

ABSTRACT

PURPOSE: To supplement the heuristics judgement of significance by taking the notice of a distribution situation in a document data base as against an arbitrary word in a document and statistically and objectively judging significance as a keyword.

CONSTITUTION: A candidate input part 11 inputs the **word** included in the **document** being a **keyword** extraction object as a candidate **word** . A **document group** 13 prepares the set of the **documents** and a **frequency** calculation part 12 calculates the **frequency** of the candidate **word** , which is the number of the **documents** including the candidate word, in the **document** group 13. An efficiency calculation part 14 calculates the efficiency of retrieval, which is efficiency for narrowing the words to the less documents when the candidate words are used as retrieval keys. A recall ability calculation part 15 calculates the recall ability of the word, which is the recall easiness of the word, on the candidate words. A significance calculation part 16 calculates the significance of the candidate words from the efficiency of retrieval and the recall ability of the word. A judgement part 17 judges the keyword appropriate degree of the candidate word from the magnitude of the significance of the candidate words.

**14/5/45 (Item 45 from file: 347)**
DIALOG(R)File 347:JAPIO

02891334 **Image available**
AUTOMATIC DOCUMENT SORTING DEVICE

PUB. NO.: 01-188934 [JP 1188934 A]
PUBLISHED: July 28, 1989 ( **19890728)**
INVENTOR(s): TAMURA ATSUSHI
APPLICANT(s): NEC CORP [000423] (A Japanese Company or Corporation), JP (Japan)
APPL. NO.: 63-013063 [JP 8813063]
FILED: January 22, 1988 (19880122)
INTL CLASS: [4] **G06F-007/28**
JAPIO CLASS: 45.1 (INFORMATION PROCESSING -- Arithmetic Sequence Units); 45.2 (INFORMATION PROCESSING -- Memory Units)
JOURNAL: Section: P, Section No. 951, Vol. 13, No. 478, Pg. 67, October 30, 1989 (19891030)

ABSTRACT

PURPOSE: To effectively sort **documents** by checking a sample **document group** to obtain the **appearing frequency information** on the key **words** of each field and knowing a key **word** having the high identifying power as well as the degree of this identifying degree.

CONSTITUTION: In a · preparatory process a key word is extracted by an automatic key word extracting means 2 for a sample document. Then the appearing frequency of the extracted key word is counted by a positive score table production means 71 for acquisition of the squared value. Then a key word having high identifying power is selected and at the same time the score of the key word showing the degrees of contribution to each field is calculated from said squared value. These calculated scores are stored in a score table storing means 8. In a field process, the means 2 ejects the key word to the document received from a document input means 1. Then the score of the key word is read out by reference to the means 8 and added to each field. The sorting operation is carried out to the fixed area of a document from its head toward a field showing the highest score

```
 14/5/58     (Item 6 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005  Thomson Derwent. All rts. reserv.

012720831     **Image available**
WPI Acc No: 1999-526943/ 199944
XRPX Acc No: N99-390310
  On-Internet information collection method
Patent Assignee: HYPERTAK INC (HYPE-N)
Inventor: CHIANG J Y; ONOE T
Number of Countries: 001  Number of Patents: 001
Patent Family:
Patent No      Kind   Date     Applicat No     Kind   Date      Week
US 5951642      A   19990914  US 97907237      A   19970806  199944  B

Priority Applications (No Type Date): US 97907237 A 19970806
Patent Details:
Patent No  Kind Lan Pg   Main IPC    Filing Notes
US 5951642    A     19 G06F-013/00

Abstract (Basic): US 5951642 A
        NOVELTY - The viewing information is acquired, when the information
    collection client program is activated for any one of the information
    viewers to view information of any one of information providers. The
    acquired viewing information is processed statistically based on which
    detailed on-Internet information onto server (3) of information
    collector is acquired.
        DETAILED DESCRIPTION - The information of the information provider
    is information at WWW, electronic mail, mailing lists and net-news of
    the information provider. The uniform resource locator (URL) of the
    WWW, time, titles, senders and dates are acquired as viewing
    information  of the  information  viewer. The acquired viewing
    information  is processed statistically in  terms  of access time,
    access  frequencies  and viewer's genders, age  groups  and regions.
    INDEPENDENT CLAIMS are also included for the following:
        (a) on-Internet information collection system;
        (b) storage medium for storing information collection client
    program for collecting information
        USE - For collecting detailed on-Internet at WWW site information
    on Internet connected to Intranet. Also for searching general files of
    information and services such as Gopher, file transfer, remote log-in.
        ADVANTAGE - Makes possible to automatically acquire the detailed
    on-internet information, thereby improves operation efficiency of the
    information provider. It is possible to become aware of things such as
    to what extent the people who are accessing his page are accessing the
    pages of which other companies while evaluating user's own web site,
    thereby information viewers constitute group of people who strongly
    reflect the market trends. The information provider who provides
    services of electronic mail, mailing lists, net-news can know more
    detailed information on his competitors, his advantages- disadvantages
    and his weakness-strength in comparison with the competitors, thereby
    improves work efficiency.
        DESCRIPTION OF DRAWING(S) - The figure depicts schematic diagram of
    on-Internet network system.
        Server of the information collector (3)
        pp; 19 DwgNo 1/12
Title Terms: INFORMATION; COLLECT; METHOD
Derwent Class: T01
International Patent Class (Main):  G06F-013/00
File Segment: EPI


 14/5/61     (Item 9 from file: 350)
```

```
012438022     **Image available**
WPI Acc No: 1999-244130/ 199920
Related WPI Acc No: 1999-045091
XRPX Acc No: N99-181663
  collection selection relative to a set of databases to obtain consistent
  relative-ranking collection selection results each iteration
Patent Assignee: INFOSEEK CORP (INFO-N)
Inventor: CHANG W I; KIRSCH S T
Number of Countries: 081  Number of Patents: 004
Patent Family:
Patent No      Kind   Date     Applicat No     Kind   Date      Week
WO 9914691      A1  19990325  WO 98US18844      A   19980910  199920  B
AU 9892282      A   19990405  AU 9892282        A   19980910  199933
US 5983216      A   19991109  US 97928294       A   19970912  199954
US 6018733      A   20000125  US 97928543       A   19970912  200012

Priority Applications (No Type Date): US 97928543 A 19970912; US 97928294 A
  19970912; US 97928542 A 19970912
Patent Details:
Patent No  Kind Lan Pg   Main IPC     Filing Notes
WO 9914691     A1 E  46 G06F-017/30
   Designated States (National): AL AM AT AU AZ BA BB BG BR BY CA CH CN CU
   CZ DE DK EE ES FI GB GE GH GM HU ID IL IS JP KE KG KP KR KZ LC LK LR LS
   LT LU LV MD MG MK MN MW MX NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR
   TT UA UG UZ VN YU ZW
   Designated States (Regional): AT BE CH CY DE DK EA ES FI FR GB GH GM GR
   IE IT KE LS LU MC MW NL OA PT SD SE SZ UG ZW
AU 9892282     A          G06F-017/30   Based on patent WO 9914691
US 6018733     A          G06F-017/30
US 5983216     A          G06F-017/30
```

Abstract (Basic): WO 9914691 A1

NOVELTY - A **collection** selection query including a set of set search **terms** is obtained. An inverse **collection** **frequency** is determined for each search **term** with respect to each database and the set of databases. A **document** frequency is determined for each search term with respect to each database. A ranking value is determined for each database based on a sum of the products of the inverse **collection** **frequencies** for the search **terms** and the **document** **frequencies** for respective search **terms** . A subset of the set of databases is selected based on set criteria dependent on the ranking value for each database.

DETAILED DESCRIPTION - The method involves: a) obtaining a **collection** selection query including a set of set search **terms** , b) determining an inverse **collection** **frequency** for each search **term** with respect to each database and the set of databases, and determining a **document** frequency for each search term with respect to each database, c) determining a ranking value for each database based on a sum of the products of the inverse **collection** **frequencies** for the search **terms** and the **document** **frequencies** for respective search **terms** , d) selecting a subset of the set of databases based on set criteria dependent on the ranking value for each database, and e) selectively repeating portions of the steps (b) through (d) with respect to each search term for each iteration of the method.

USE - The method is used to permit iterative performance of collection selection relative to a set of databases, where each database includes several documents, to obtain consistent relative-ranking collection selection results each iteration.

ADVANTAGE - Improves selection of most relevant collections for searching based on an ad hoc query.

DESCRIPTION OF DRAWING(S) - The drawing shows a flow diagram illustrating the operation in supporting a meta-index database

```
    construction and user search.
        pp; 46 DwgNo 1/6
Title Terms: COLLECT; SELECT; RELATIVE; SET; OBTAIN; CONSISTENT; RELATIVE;
  RANK; COLLECT; SELECT; RESULT; ITERATIVE
Derwent Class: T01; W01
International Patent Class (Main): G06F-017/30
File Segment: EPI


 14/5/62     (Item 10 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005  Thomson Derwent. All rts. reserv.

012012988     **Image available**
WPI Acc No: 1998-429898/ 199837
Related WPI Acc No: 2005-032906
XRPX Acc No: N98-335702
  Document retrieval appts for retrieving desired documents stored in
  computer on network e.g. internet - calculates degree of similarity by
  weighting with structure of document and occurrence frequency of keyword
  in document
Patent Assignee: KOKUSAI DENSHIN DENWA CO LTD (KOKU  ); DAINI DENDEN KK
  (DAIN-N)
Inventor: AOKI K; HASHIMOTO K; MATSUMOTO K
Number of Countries: 026  Number of Patents: 006
Patent Family:
Patent No      Kind   Date     Applicat No     Kind   Date      Week
EP 859330       A1   19980819  EP 98301003      A   19980211   199837  B
JP 10222534     A    19980821  JP 9741429       A   19970212   199844
JP 10254905     A    19980925  JP 9767496       A   19970306   199849
US 6078913      A    20000620  US 9822280       A   19980211   200035
JP 3632359      B2   20050323  JP 9767496       A   19970306   200522
JP 3632354      B2   20050323  JP 9741429       A   19970212   200522
Priority Applications (No Type Date): JP 9767496 A 19970306; JP 9741429 A
  19970212
Patent Details:
Patent No  Kind Lan Pg   Main IPC    Filing Notes
EP 859330     A1 E  23 G06F-017/30
   Designated States (Regional): AL AT BE CH DE DK ES FI FR GB GR IE IT LI
   LT LU LV MC MK NL PT RO SE SI
JP 10222534   A      7 G06F-017/30
JP 10254905   A     11 G06F-017/30
US 6078913    A        G06F-017/21
JP 3632359    B2    12 G06F-017/30   Previous Publ. patent JP 10254905
JP 3632354    B2     8 G06F-017/30   Previous Publ. patent JP 10222534

Abstract (Basic): EP 859330 A
        The appts includes a cluster database storing a cluster of a number
    of node information linked for clustering the documents to a
    hierarchical tree structure based on degree of similarity in all
    documents. The node information has the posted end addresses to be
    posted when the documents positioned to the lower layer of the node
    information is updated. The control device is posted to the posted end
    address in the node information encountered on the way to follow links
    of the cluster by device of the cluster database when the document is
    updated. The degree of similarity is calculated by weighting with the
    structure of the  document  and the  occurrence  frequency  of a
    keyword  in the  document . The  cluster  is executed by linking the
    similar  documents  closely with each other based on degree of
    similarity.
        Dwg.1b/9
Title Terms: DOCUMENT; RETRIEVAL; APPARATUS; RETRIEVAL; DOCUMENT; STORAGE;
  COMPUTER; NETWORK; CALCULATE; DEGREE; SIMILAR; WEIGHT; STRUCTURE;
  DOCUMENT; OCCUR; FREQUENCY; KEYWORD; DOCUMENT
Derwent Class: T01
```

```
International Patent Class (Main): G06F-017/21 ; G06F-017/30
File Segment: EPI


 14/5/65     (Item 13 from file: 350)
DIALOG(R)File 350:Derwent WPIX
(c) 2005  Thomson Derwent. All rts. reserv.

011738769     **Image available**
WPI Acc No: 1998-155679/ 199814
XRPX Acc No: N98-124305
  Document classification apparatus for hypertext in internet - forms first
  stage  document  cluster , by grouping various  information  such as
  link relations and  appearance  frequency  of matching  word  in stored
  documents
Patent Assignee: FUJI XEROX CO LTD (XERF  )
Number of Countries: 001  Number of Patents: 001
Patent Family:
Patent No      Kind   Date     Applicat No     Kind   Date      Week
JP 10027125     A   19980127  JP 96199543      A   19960711  199814  B

Priority Applications (No Type Date): JP 96199543 A 19960711
Patent Details:
Patent No  Kind Lan Pg   Main IPC    Filing Notes
JP 10027125   A     10 G06F-012/00

Abstract (Basic): JP 10027125 A
        The apparatus has a memory (11) which stores some documents to be
    processed. The link relation between the documents, is stored in a link
    relation memory (12). The frequency of appearance of word in each
    document is stored in a matching information memory (13).
        The individually stored information are then grouped to form a
    first stage document cluster and cluster analysis is carried out. A
    classifier classifies the stored documents and the classified result is
    output suitably.
        ADVANTAGE - Performs classification using corresponding links,
    accurately.
        Dwg.1/13
Title Terms: DOCUMENT; CLASSIFY; APPARATUS; FORM; FIRST; STAGE; DOCUMENT;
  CLUSTER; GROUP; VARIOUS; INFORMATION; LINK; RELATED; APPEAR; FREQUENCY;
  MATCH; WORD; STORAGE; DOCUMENT
Derwent Class: T01
International Patent Class (Main): G06F-012/00
International Patent Class (Additional): G06F-017/27 ; G06F-017/30
File Segment: EPI
File   8:Ei Compendex(R) 1970-2005/Jul W5
         (c) 2005 Elsevier Eng.  Info. Inc.
File  35:Dissertation Abs Online 1861-2005/Jul
         (c) 2005 ProQuest Info&Learning
File  65:Inside Conferences 1993-2005/Aug W1
         (c) 2005 BLDSC all rts. reserv.
File   2:INSPEC 1969-2005/Jul W5
         (c) 2005 Institution of Electrical Engineers
File  94:JICST-EPlus 1985-2005/Jun W3
         (c)2005 Japan Science and Tech Corp(JST)
File   6:NTIS 1964-2005/Jul W5
         (c) 2005 NTIS, Intl Cpyrght All Rights Res
File 144:Pascal 1973-2005/Jul W5
         (c) 2005 INIST/CNRS
File 434:SciSearch(R) Cited Ref Sci 1974-1989/Dec
         (c) 1998 Inst for Sci Info
File  34:SciSearch(R) Cited Ref Sci 1990-2005/Jul W5
         (c) 2005 Inst for Sci Info
File  99:Wilson Appl. Sci & Tech Abs 1983-2005/Jul
         (c) 2005 The HW Wilson Co.
File 266:FEDRIP 2005/Jun
         Comp & dist by NTIS, Intl Copyright All Rights Res
File  95:TEME-Technology & Management 1989-2005/Jul W1
         (c) 2005 FIZ TECHNIK
File 583:Gale Group Globalbase(TM) 1986-2002/Dec 13
         (c) 2002 The Gale Group
File 438:Library Lit. & Info. Science 1984-2005/Jul
         (c) 2005 The HW Wilson Co

Set     Items   Description
S1    9854210   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
                 OR CATEGORIES OR CLASS OR CLASSES OR GROUP? ? OR CLUSTER? ? -
                OR FAMILY OR FAMILIES OR COLLECTION? ? OR DOMAIN? ?
S2    6498890   FREQUEN???? OR OCCURR? OR INCIDENCE? ? OR HOW()OFTEN OR AP-
                PEAR?
S3       2636   S2(5N)(TERM? ? OR WORD? ? OR KEYWORD? ? OR ELEMENT? ?)(5N)-
                (DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? OR RECORD?
                 ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMATION OR CON-
                TENT? ?)(5N)S1
S4    3689710   RATIO? ? OR PERCENT???? OR PROPORTION??
S5        242   S3 AND S4
S6        193   RD (unique items)
S7        135   S6 NOT PY=2000:2005
S8      96239   (FREQUEN???? OR OCCURR? OR INCIDENCE? ? OR HOW()OFTEN OR A-
                PPEAR?)(5N)(TERM? ? OR WORD? ? OR KEYWORD? ? OR ELEMENT? ?)
S9         72   S7 AND S8
S10       813   S8(5N)(DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? -
                OR RECORD? ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMAT-
                ION OR CONTENT? ?)(5N)S1
S11        59   S10 AND S4
S12        46   RD (unique items)
S13        30   S12 NOT PY=2000:2005
S14   2139336   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
                 OR CATEGORIES
S15       217   S8(5N)(DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? -
                OR RECORD? ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMAT-
                ION OR CONTENT? ?)(5N)S14
S16       180   RD (unique items)
S17       124   S16 NOT PY=2000:2005
S18       116   S17 NOT S13
```

**18/5/1 (Item 1 from file: 8)**
DIALOG(R)File 8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

05099284 E.I. No: EIP98084344904
**Title: Topic extraction with multiple topic-words in broadcast-news speech**
Author: Ohtsuki, K.; Matsutoka, T.; Matsunaga, S.; Furui, S.
Corporate Source: NTT Human Interface Lab, Kanagawa, Jpn
Conference Title: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP. Part 1 (of 6)
Conference Location: Seattler, WA, USA Conference Date: 19980512-19980515
Sponsor: IEEE
E.I. Conference No.: 48801
Source: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings v 1 1998. IEEE, Piscataway, NJ, USA,98CH36181. p 329-332
Publication Year: 1998
CODEN: IPRODJ ISSN: 0736-7791
Language: English
Document Type: CA; (Conference Article) Treatment: T; (Theoretical); X; (Experimental)
Journal Announcement: 9810W3
Abstract: This paper reports on topic extraction in Japanese broadcast-news speech. We studied, using continuous speech recognition, the extraction of several topic-words from broadcast-news. A combination of multiple topic-words represents the content of the news. This is a more detailed and more flexible approach than using a single word or a single category. A topic-extraction model shows the degree of relevance between each topic-word and each word in the article. For all words in an article, topic-words which have high total relevance score are extracted. We trained the **topic** -extraction model with five years of newspapers, using the **frequency** of **topic - words** taken from headlines and words in **articles** . The degree of relevance between topic-words and words in articles is calculated on the basis of statistical measures, i.e., mutual information or the chi **2-value. In topic-extraction experiments for recognized broadcast-news speech, we extracted five topic-words from the 10-best hypotheses using a chi **2-based model and found that 76.6% of them agreed with the topic-words chosen by subjects. (Author abstract) 13 Refs.
Descriptors: *Continuous speech recognition; Mathematical models; Information retrieval systems; Statistical methods; Probability; Markov processes; Linguistics
Identifiers: Topic extraction model; Degree of relevance; Hidden Markov models; Language modelling
Classification Codes:
751.5 (Speech); 903.3 (Information Retrieval & Use); 922.2 (Mathematical Statistics); 922.1 (Probability Theory)
751 (Acoustics); 921 (Applied Mathematics); 903 (Information Science); 922 (Statistical Methods)
75 (ACOUSTICAL TECHNOLOGY); 92 (ENGINEERING MATHEMATICS); 90 (GENERAL ENGINEERING)

**18/5/8 (Item 8 from file: 8)**
DIALOG(R)File 8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

03695170 E.I. No: EIP93081064122
**Title:** Subject **relationship between** articles **determined by co-**occurrence **of** keywords **in citing and cited titles**
Author: Ali, S. Nazim
Corporate Source: Univ of Bahrain, Isa Town, Bahrain
Source: Journal of Information Science: Principles & Practice (Amsterdam) v 19 n 3 1993. p 225-232

Publication Year: 1993
CODEN: JISCDI ISSN: 0165-5515
Language: English
Document Type: JA; (Journal Article) Treatment: A; (Applications); G; (General Review)
Journal Announcement: 9310W4

Abstract: It is assumed that a paper which cites an earlier document shares a subject relationship with that particular document. In order to determine if this assumption is valid, a study was conducted by analyzing 1000 articles from the Science Citation Index and Social Sciences Citation Index. These articles were selected in ten different disciplines by using a purposive sampling technique. Various Spearman's Correlation Coefficient tests were computed to find out if a subject relationship existed between the Articles which have the same keywords in their titles (Parent Articles and Related Records). Through the analysis the hypothesis has been verified showing that there is a relationship between the articles which are citing the same references. This was determined by co-occurrences of the same keywords among the shared references. However, there are some unique differences in the science and the social science disciplines that exist in these two databases. (Author abstract) 14 Refs.

Descriptors: *Indexing (of information); Information retrieval; Vocabulary control; Information analysis; Database systems; Correlation methods; Classification (of information); Bibliographic retrieval systems; Social sciences; Information management

Identifiers: Science citation index; Social sciences citation index; Spearman's correlation coefficient

Classification Codes:
903.1 (Information Sources & Analysis); 903.3 (Information Retrieval & Use); 723.3 (Database Systems); 912.2 (Management)
903 (Information Science); 723 (Computer Software); 921 (Applied Mathematics); 912 (Industrial Engineering & Management)
90 (GENERAL ENGINEERING); 72 (COMPUTERS & DATA PROCESSING); 92 (ENGINEERING MATHEMATICS); 91 (ENGINEERING MANAGEMENT)

**18/5/14 (Item 14 from file: 8)**
DIALOG(R)File 8:Ei Compendex(R)
(c) 2005 Elsevier Eng. Info. Inc. All rts. reserv.

01634948 E.I. Monthly No: EIM8403-018958
**Title: FUZZY MEASURE OF AGREEMENT BETWEEN MACHINE AND MANUAL ASSIGNMENT OF DOCUMENTS TO SUBJECT CATEGORIES.**
Author: Cerny, Barbara A.; Okseniuk, Anna; Lawrence, J. Dennis
Corporate Source: Dialog Information Services Inc, Palo Alto, Calif, USA
Conference Title: Productivity in the Information Age, Proceedings of the 46th ASIS Annual Meeting.
Conference Location: Washington, DC, USA Conference Date: 19831002
Sponsor: ASIS, Washington, DC, USA
E.I. Conference No.: 03341
Source: Proceedings of the ASIS Annual Meeting 46th v 20 1983. Publ for ASIS by Knowledge Industry Publ, White Plains, NY, USA p 265
Publication Year: 1983
CODEN: PAISDQ ISSN: 0044-7870 ISBN: 0-86729-072-2
Language: English
Document Type: PA; (Conference Paper)
Journal Announcement: 8403
Descriptors: *INFORMATION SCIENCE--*Indexing
Identifiers: FUZZY MEASURE; AGREEMENT BETWEEN MACHINE AND MANUAL ASSIGNMENT OF **DOCUMENTS** ; **SUBJECT CATEGORIES** ; MULTIPLE **SUBJECT CATEGORIES** ; WORD STEM **FREQUENCY** ; AUTOMATIC ASSIGNMENT BY **WORD FREQUENCY** ANALYSIS OF ABSTRACTS; TRAINING SET; FUZZY PREDICTION PROBLEM
Classification Codes:
901 (Engineering Profession)
90 (GENERAL ENGINEERING)

**18/5/53 (Item 1 from file: 2)**
DIALOG(R)File 2:INSPEC

6537919 INSPEC Abstract Number: C2000-04-6130D-018
**Title: News article classification based on categorical points from keywords in backdata**
Author(s): Jo, T.C.
Author Affiliation: Inf. R&D Center, Samsung SDS, Seoul, South Korea
Conference Title: Computational Intelligence for Modelling, Control and Automation. Intelligent Image Processing, Data Analysis and Information Retrieval (Concurrent Systems Engineering Series Vol.56) p.211-14
Editor(s): Mohammadian, M.
Publisher: IOS Press, Amsterdam, Netherlands
Publication Date: 1999 Country of Publication: Netherlands xi+338 pp.
ISBN: 90 5199 475 3 Material Identity Number: XX-2000-00480
Conference Title: Computational Intelligence for Modelling, Control and Automation. Intelligent Image Processing, Data Analysis and Information Retrieval
Conference Date: 17-19 Feb. 1999 Conference Location: Vienna, Austria
Language: English Document Type: Conference Paper (PA)
Treatment: Practical (P)
Abstract: A scheme of automatic document classification is presented. Previously, documents have been classified according to their contents manually. Therefore, it is very costly to assign a category to them because a human investigates their contents. As the amount of data stored in storage media is increased exponentially, it becomes necessary to store documents according to their category, to access them easily. Automatic text classification is needed to store documents like that. Before performing text classification, back **data** should be constructed. The back **data** stores the **information** about **keywords** : the **frequency** for each **category** , the number of **documents** for each **category** . A **document** is represented with a list of keywords. Categorical points to each **category** are computed by summing the **frequency** of each **keyword** from back **data** , or the number of **documents** from it. The **category** that contains the largest categorical points is selected as the category of a document. In the results of an experiment with news article classification, precision is about 98%. (11 Refs)
Subfile: C
Descriptors: classification; publishing; text analysis; word processing
Identifiers: news article classification; categorical points; backdata keywords; automatic document classification; storage media; automatic text classification
Class Codes: C6130D (Document processing techniques); C7240 (Information analysis and indexing); C7230 (Publishing and reproduction)

**18/5/62 (Item 10 from file: 2)**
DIALOG(R)File 2:INSPEC

04329853 INSPEC Abstract Number: C9303-7240-007
**Title: An automatic document classification method based on a semantic category frequency analysis**
Author(s): Kawai, A.
Author Affiliation: NTT Commun. & Inf. Processing Lab., Ibaraki, Japan
Journal: Transactions of the Information Processing Society of Japan vol.33, no.9 p.1114-22
Publication Date: 1992 Country of Publication: Japan
CODEN: JSGRD5 ISSN: 0387-5806
Language: Japanese Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: Describes the execution image of the **document** classification

system; the hierarchy of a semantic **category** (a general noun); semantic **frequency** tables; **words** and **categories** selected from each division; recall/precision graphs of classification systems and the evaluation of reference dictionaries; and the relation between classification fields and semantic categories. (17 Refs)
Subfile: C
Descriptors: classification; information retrieval
Identifiers: word selection; automatic document classification method; semantic category frequency analysis; execution image; noun; recall/precision graphs; reference dictionaries; classification fields
Class Codes: C7240 (Information analysis and indexing)

**18/5/66 (Item 14 from file: 2)**
DIALOG(R)File 2:INSPEC
(c) 2005 Institution of Electrical Engineers. All rts. reserv.

03488821 INSPEC Abstract Number: C89068881
**Title: Automated indexing for making of a newspaper article database**
Author(s): Kamio, T.
Author Affiliation: Nihon Keizai Shimbun Inc., Tokyo, Japan
Journal: Joho Kanri vol.32, no.4 p.283-93
Publication Date: July 1989 Country of Publication: Japan
CODEN: JOKAAB ISSN: 0021-7298
Language: Japanese Document Type: Journal Paper (JP)
Treatment: Practical (P)
Abstract: Automated indexing has been widely employed in the process of making newspaper article databases. It is essential to speed up the compiling time of the said databases as a large amount of articles are produced daily, and to conserve manpower with the aid of computers. However, indexed terms which are extracted by current automated indexing systems have no links with subject analysis so they are not considered to be keywords in a strict sense. Thus, the system of Nihon Keizai Shimbun KK enables the justification of certain keywords to a certain extent, based on the two clues: which location the extracted **term occurred** and whether or not the **subject** area of the **article** corresponds to the thesaurus class of the extracted term by using characteristics peculiar to newspaper articles. An experiment involving assigning keywords which have not occurred in articles was also conducted and a fairly good result was obtained. (9 Refs)
Subfile: C
Descriptors: indexing; information services; thesauri
Identifiers: newspaper article databases; compiling time; indexed terms; automated indexing systems; subject analysis; Nihon Keizai Shimbun KK; extracted term; subject area; thesaurus class; keywords
Class Codes: C7240 (Information analysis and indexing); C7210 ( Information services and centres)

**18/5/79 (Item 7 from file: 94)**
DIALOG(R)File 94:JICST-EPlus
(c)2005 Japan Science and Tech Corp(JST). All rts. reserv.

02224924 JICST ACCESSION NUMBER: 94A0929607 FILE SEGMENT: JICST-E
**Document Classification Using Important Kanji Characters Extracted by x2 Method.**
WATANABE YASUHIKO (1); TAKEUCHI MASAHITO (1); MURATA MASAKI (1); NAGAO MAKOTO (1)
(1) Kyoto Univ., Fac. of Eng.
Denshi Joho Tsushin Gakkai Gijutsu Kenkyu Hokoku(IEIC Technical Report (Institute of Electronics, Information and Communication Enginners), 1994, VOL.94,NO.292(NLC94 22-25v27-31), PAGE.23-30, FIG.2, TBL.5, REF.6
JOURNAL NUMBER: S0532BBG
UNIVERSAL DECIMAL CLASSIFICATION: 002.5:025 681.3:80
LANGUAGE: Japanese COUNTRY OF PUBLICATION: Japan

DOCUMENT TYPE: Journal
ARTICLE TYPE: Original paper
MEDIA TYPE: Printed Publication

ABSTRACT: It is generally recognized to classify a given **document** into several **categories** by using technical **words** which preferably **appear** in one **category** than the other. However, we have much difficulties to extract the technical words properly from Japanese sentences. Instead of these technical words, we adopted kanji characters which preferably appear in one category than the other. In this paper, we describe how to extract the important kanji characters for document classification by x2 method and how to classfy documents in a simple pattern classification method. Then, we examined our method and the correct recognition scores for "TENSEI JINGO", editorial articles, and articles in "SCIENCE" were 41.6%, 77.9%, and 92.7%, respectively. (author abst.)

```
File 348:EUROPEAN PATENTS 1978-2005/Jul W05
         (c) 2005 European Patent Office
File 349:PCT FULLTEXT 1979-2005/UB=20050804,UT=20050728
         (c) 2005 WIPO/Univentio

Set     Items   Description
S1    1051996   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
                 OR CATEGORIES OR CLASS OR CLASSES OR GROUP? ? OR CLUSTER? ? -
                OR FAMILY OR FAMILIES OR COLLECTION? ? OR DOMAIN? ?
S2      61214    (FREQUEN???? OR OCCURR? OR INCIDENCE? ? OR HOW()OFTEN OR A-
                PPEAR?)(5N)(TERM? ? OR WORD? ? OR KEYWORD? ? OR ELEMENT? ?)
S3        546   S2(5N)(DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? -
                 OR RECORD? ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMAT-
                ION OR CONTENT? ?)(5N)S1
S4     793146   RATIO? ? OR PERCENT???? OR PROPORTION??
S5         22   S3(50N)S4
S6         15   S5 AND IC=G06F
S7         10   S6 AND AY=(1970:1999)/PR
S8         10   S6 AND AY=1970:1999
S9         10   S7:S8
S10       524   S3 NOT S5
S11       334   S10 AND IC=G06F
S12    452704   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
                 OR CATEGORIES
S13       170   S2(5N)(DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? -
                OR RECORD? ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMAT-
                ION OR CONTENT? ?)(5N)S12
S14       132   S13 AND IC=G06F
S15       123   S14 NOT S5
S16        67   S15 AND AC=US/PR
S17        30   S16 AND AY=(1970:1999)/PR
S18        30   S15 AND PY=1970:1999
S19        45   S17:S18
S20        45   IDPAT (sorted in duplicate/non-duplicate order)
```

9/5/9     (Item 3 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT

00483339    **Image available**

**METHODS FOR ITERATIVELY AND INTERACTIVELY PERFORMING COLLECTION SELECTION IN FULL TEXT SEARCHES**

**PROCEDES PERMETTANT D'EFFECTUER UNE SELECTION DE COLLECTIONS DANS DES RECHERCHES SUR TEXTE INTEGRAL**

Patent Applicant/Assignee:
  INFOSEEK CORPORATION,
Inventor(s):
  KIRSCH Steven T,
  CHANG William I,
Patent and Priority Information (Country, Number, Date):
  Patent:              WO 9914691 A1 19990325
  Application:         WO 98US18844 19980910  (PCT/WO US9818844)
  Priority Application: US 97928542 19970912; US 97928543 19970912; US 97928294 19970912
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)

```
  AL AM AT AU AZ BA BB BG BR BY CA CH CN CU CZ DE DK EE ES FI GB GE GH GM
  HU ID IL IS JP KE KG KP KR KZ LC LK LR LS LT LU LV MD MG MK MN MW MX NO
  NZ PL PT RO RU SD SE SG SI SK SL TJ TM TR TT UA UG UZ VN YU ZW GH GM KE
  LS MW SD SZ UG ZW AM AZ BY KG KZ MD RU TJ TM AT BE CH CY DE DK ES FI FR
  GB GR IE IT LU MC NL PT SE BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
```

Main International Patent Class: **G06F-017/30**
Publication Language: English
Fulltext Availability:
  Detailed Description
  Claims
Fulltext Word Count: 11731

English Abstract

A method of selecting a subset of a plurality of document collections for searching in response to a predetermined query is based on accessing a meta-information data file that describes the query significant search terms that are present in a particular document collection correlated to normalized document usage frequencies of such terms within the documents of each document collection. By access to the meta-information data file, a relevance score for each of the document collections is determined. The method then returns an identification of the subset of the plurality of document collections having the highest relevance scores for use in evaluating the predetermined query. The meta-information data file may be constructed to include document normalized term frequencies and other contextual information that can be evaluated in the application of a query against a particular document collection.

15 A method of selecting a subset of a set of document collections containing documents to search based upon a predetermined query text including a
search term, said method comprising the steps of
a) accessing a meta-file representative of said set of **document** collections, including a search term occurrence list;
b) determining a document frequency term for said search term relative to each of said document collections within said set of document collections and an inverse collection **frequency term** for said set of document collections, said inverse **collection frequency** term being proportional to a ratio of the number of **documents** in 0 said set of document collections and the number of documents in set of document 1 collections that include said search term;
c) determining a term ranking for each of said **document collections** 3 that is proportional to the respective said **document frequency terms** and said inverse
**collection frequency term** ;
d) selecting said subset of said set of **document collections** based on the 6 relative term ranking of each of said **document** collections.

```
20/3,K/2     (Item 2 from file: 349)
DIALOG(R)File 349:PCT FULLTEXT
(c) 2005 WIPO/Univentio. All rts. reserv.

00507930
SCORING OF TEXT UNITS
DENOMBREMENT D'UNITES DE TEXTE
Patent Applicant/Assignee:
  SHARP KABUSHIKI KAISHA,
  SANFILIPPO Antonio Pietro,
Inventor(s):
  SANFILIPPO Antonio Pietro,
Patent and Priority Information (Country, Number, Date):
  Patent:               WO 9939282 A1  19990805
  Application:          WO 99JP259 19990122  (PCT/WO JP9900259)
  Priority Application: GB 981784 19980129
Designated States:
(Protection type is "patent" unless otherwise stated - for applications
prior to 2004)
  CA CN IN JP US AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE
Publication Language: English
Fulltext Word Count: 8133

Patent and Priority Information (Country, Number, Date):
  Patent:                ... 19990805
Main International Patent Class:  G06F-017/30
Fulltext Availability:
  Claims
Publication Year:  1999

Claim
...  operating on a text as claimed in claim 4,
  which comprises the further step of keeping a  record  of
  the  word  spelling associated with each  occurrence  of a
   subject  code in a text unit, and wherein during said
  summing step occurrences of the same subject code...


 20/3,K/6     (Item 6 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS
(c) 2005 European Patent Office. All rts. reserv.

00929211
A data processing method and apparatus for indentifying a classification to
    which data belongs
Ein  Datenverarbeitungsverfahren  zum Ermitteln der zu den Daten gehorenden
    Klassifikation
Methode  de  traitement de donnees pour identifier la propre classification
    des donnees
PATENT ASSIGNEE:
  CANON KABUSHIKI KAISHA, (542361), 30-2, 3-chome, Shimomaruko, Ohta-ku,
    Tokyo, (JP), (Proprietor designated states: all)
INVENTOR:
  Elworthy, David, c/o Canon Res. CTR. Europe Ltd., 1 Occam Court, Occam
    Road, Surrey Research Park, Guildford, Surrey GU2 5YJ, (GB)
LEGAL REPRESENTATIVE:
  Beresford, Keith Denis Lewis et al (28273), BERESFORD & Co. 2-5 Warwick
    Court, High Holborn, London WC1R 5DH, (GB)
PATENT (CC, No, Kind, Date):  EP 847018  A1  980610 (Basic)
                              EP 847018  B1  021113
APPLICATION (CC, No, Date):   EP 97309691 971202;
PRIORITY (CC, No, Date): GB 9625284 961204
DESIGNATED STATES: FR; GB; IT
INTERNATIONAL PATENT CLASS:  G06F-017/27 ; G06K-009/68
ABSTRACT WORD COUNT: 110
```

NOTE:
  Figure number on first page: 3

LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:

| Available Text | Language | Update | Word Count |
|---|---|---|---|
| CLAIMS A | (English) | 199824 | 3141 |
| CLAIMS B | (English) | 200246 | 2743 |
| CLAIMS B | (German) | 200246 | 2555 |
| CLAIMS B | (French) | 200246 | 2959 |
| SPEC A | (English) | 199824 | 8280 |
| SPEC B | (English) | 200246 | 8007 |
| Total word count - document A | | | 11424 |
| Total word count - document B | | | 16264 |
| Total word count - documents A + B | | | 27688 |

INTERNATIONAL PATENT CLASS: **G06F-017/27** ...

...SPECIFICATION e.g. legal or scientific. The type of document can be identified from the layout of the **document** e.g. the position and/or shape of the paragraphs. The **topic** of a **document** can be identified by identifying the **occurrence** of certain **words** within the document and comparing these with the probability of the occurrence of these documents in various...

...SPECIFICATION e.g. legal or scientific. The type of document can be identified from the layout of the **document** e.g. the position and/or shape of the paragraphs. The **topic** of a **document** can be identified by identifying the **occurrence** of certain **words** within the document and comparing these with the probability of the occurrence of these documents in various...

**20/3,K/14 (Item 14 from file: 348)**
DIALOG(R)File 348:EUROPEAN PATENTS

00656152
**Method of processing a corpus of electronically stored documents**
**Verfahren zur Verarbeitung mehrerer elektronisch gespeicherte Dokumente**
**Procede pour traiter plusieurs documents stockes electroniquement**
PATENT ASSIGNEE:
  XEROX CORPORATION, (219781), Xerox Square - 020, Rochester New York 14644 , (US), (Proprietor designated states: all)
INVENTOR:
  Pedersen, Jan O., 3913 Bibbits Drive, Palo Alto, California 94303, (US)
  Karger, David R., 76E Escondido Village, Stanford, California 94305, (US)
  Cutting, Douglass R., 726 Oak Grove Avenue, No. 3, Menlo Park, California 94025, (US)
LEGAL REPRESENTATIVE:
  Grunecker, Kinkeldey, Stockmair & Schwanhausser Anwaltssozietat (100721) , Maximilianstrasse 58, 80538 Munchen, (DE)
PATENT (CC, No, Kind, Date): EP 631245 A2 941228 (Basic)
                             EP 631245 A3 950222
                             EP 631245 B1 000301
APPLICATION (CC, No, Date): EP 94304471 940620;
PRIORITY (CC, No, Date): US 79292 930621
DESIGNATED STATES: DE; FR; GB
INTERNATIONAL PATENT CLASS: **G06F-017/30**
ABSTRACT WORD COUNT: 77
NOTE:
  Figure number on first page: 7

LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:

```
Available Text  Language    Update     Word Count
      CLAIMS B  (English)  200009       411
      CLAIMS B   (German)  200009       405
      CLAIMS B   (French)  200009       435
      SPEC B    (English)  200009      5342
Total word count - document A             0
Total word count - document B          6593
Total word count - documents A + B     6593

INTERNATIONAL PATENT CLASS:  G06F-017/30

...SPECIFICATION comprise suggestive text determined automatically from
  documents in each cluster. Each cluster summary includes two types of
  information : a list of  topical  words  occurring  most often in the
  documents  of the cluster, and the titles of a few typical documents in
  the cluster. The summaries are...


 20/3,K/15     (Item 15 from file: 348)
DIALOG(R)File 348:EUROPEAN PATENTS
(c) 2005 European Patent Office. All rts. reserv.

00577496
A TEXT MANAGEMENT SYSTEM
EIN TEXTVERWALTUNGSSYSTEM
SYSTEME DE GESTION DE TEXTE
PATENT ASSIGNEE:
  WANG LABORATORIES INC., (333560), One Industrial Avenue, Lowell, MA 01851
    , (US), (Proprietor designated states: all)
INVENTOR:
  KADASHEVICH, Julie, A., 43 Sherburne Avenue, Tyngsboro, MA 01879, (US)
  HARVEY, Mary, F., 505 Summer Avenue, Reading, MA 01867, (US)
  CLARK, Cheryl, 96 Bow Street, Arlington, MA 02174, (US)
LEGAL REPRESENTATIVE:
  Behrens, Dieter, Dr.-Ing. (1701), Wuesthoff & Wuesthoff Patent- und
    Rechtsanwalte Schweigerstrasse 2, 81541 Munchen, (DE)
PATENT (CC, No, Kind, Date):  EP 592402  A1  940420 (Basic)
                              EP 592402  B1  010801
                              WO 9214214  920820
APPLICATION (CC, No, Date):   EP 91904540 910201;  WO 91US739  910201
PRIORITY (CC, No, Date): EP 91904540 910201; WO 91US739 910201
DESIGNATED STATES: BE; DE; FR; GB; NL
INTERNATIONAL PATENT CLASS:  G06F-017/27 ;  G06F-017/30
NOTE:
  No A-document published by EPO
LANGUAGE (Publication,Procedural,Application): English; English; English
FULLTEXT AVAILABILITY:
Available Text  Language    Update     Word Count
      CLAIMS B  (English)  200131      1119
      CLAIMS B   (German)  200131      1115
      CLAIMS B   (French)  200131      1172
      SPEC B    (English)  200131      9198
Total word count - document A             0
Total word count - document B         12604
Total word count - documents A + B    12604

INTERNATIONAL PATENT CLASS:  G06F-017/27 ...

... G06F-017/30

...SPECIFICATION identified by running samples of text through intelligent
  filter 104 and then analyzing the results to identify  words  that
  appear  at the output but clearly do not convey  topic   information .
  Thus, another value of stop list 106 is that it serves to catch those few
  words that...
```

**20/3,K/26 (Item 26 from file: 349)**
DIALOG(R)File 349:PCT FULLTEXT

00796201 **Image available**
**SYSTEM AND METHOD FOR LOCATION, UNDERSTANDING AND ASSIMILATION OF DIGITAL DOCUMENTS THROUGH ABSTRACT INDICIA**
**SYSTEME ET PROCEDE DE LOCALISATION, DE COMPREHENSION ET D'ASSIMILATION DE DOCUMENTS NUMERIQUES PAR DES INDICES DE RESUMES**
Patent Applicant/Inventor:
HUSSAM Ali, 1908 Walden Court, Columbia, MO 65203, US, US (Residence), -- (Nationality)
Legal Representative:
POLSTER Philip B II (agent), Polster, Lieder, Woodruff & Lucchesi, 763 South New Ballas Road, St. Louis, MO 63141, US,
Patent and Priority Information (Country, Number, Date):
Patent: WO 200129709 A1 20010426 (WO 0129709)
Application: WO 2000US29009 20001019 (PCT/WO US0029009)
Priority Application: US 99160622 19991020; US 2000178745 20000128
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
AE AG AL AM AT AU AZ BA BB BG BR BY BZ CA CH CN CR CU CZ DE DK DM DZ EE
ES FI GB GD GE GH GM HR HU ID IL IN IS JP KE KG KP KR KZ LC LK LR LS LT
LU LV MA MD MG MK MN MW MX MZ NO NZ PL PT RO RU SD SE SG SI SK SL TJ TM
TR TT TZ UA UG US UZ VN YU ZA ZW
(EP) AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE
(OA) BF BJ CF CG CI CM GA GN GW ML MR NE SN TD TG
(AP) GH GM KE LS MW MZ SD SL SZ TZ UG ZW
(EA) AM AZ BY KG KZ MD RU TJ TM
Publication Language: English
Filing Language: English
Fulltext Word Count: 24412

Main International Patent Class: **G06F-017/30**
Fulltext Availability:
Detailed Description

Detailed Description
... to find books to match a request for a topic, they first look at books with the **topic** in the title.

Search engines operate the same way. **Pages** with **keywords appearing** in the title are assumed to be more relevant to the **topic** than others. Search engines will also check to see if the **keywords appear** near the top of a web page, such as in the headline or in the first few paragraphs of text. They assume that any **page** relevant to the **topic** will mention those **words** near the beginning.

**Frequency** is another major factor in how search engines determine relevancy. A search engine will analyse how often...

**20/3,K/36 (Item 36 from file: 349)**
DIALOG(R)File 349:PCT FULLTEXT

00468838
**METHOD AND APPARATUS FOR AUTOMATICALLY IDENTIFYING KEYWORDS WITHIN A DOCUMENT**
**PROCEDE ET SYSTEME POUR IDENTIFIER AUTOMATIQUEMENT DES MOTS CLES DANS UN DOCUMENT**
Patent Applicant/Assignee:

NATIONAL RESEARCH COUNCIL OF CANADA,
Inventor(s):
TURNEY Peter D,
Patent and Priority Information (Country, Number, Date):
Patent: WO 9859303 A1 **19981230**
Application: WO 98CA469 19980514 (PCT/WO CA9800469)
Priority Application: US 97880392 19970623
Designated States:
(Protection type is "patent" unless otherwise stated - for applications prior to 2004)
JP AT BE CH CY DE DK ES FI FR GB GR IE IT LU MC NL PT SE
Publication Language: English
Fulltext Word Count: 6378

Patent and Priority Information (Country, Number, Date):
Patent: ... **19981230**
Main International Patent Class: **G06F-017/30**
Fulltext Availability:
Detailed Description
Publication Year: **1998**

Detailed Description
... indexing documents is likely required to have some knowledge of the terms and understanding of the particular **subject** matter being indexed

Listing the most **frequent words** in the **document** with the exception of stop words usually results in a relatively low-quality list of keywords, especially...

```
File 275:Gale Group Computer DB(TM) 1983-2005/Aug 09
         (c) 2005 The Gale Group
File 621:Gale Group New Prod.Annou.(R) 1985-2005/Aug 09
         (c) 2005 The Gale Group
File 636:Gale Group Newsletter DB(TM) 1987-2005/Aug 08
         (c) 2005 The Gale Group
File  16:Gale Group PROMT(R) 1990-2005/Aug 08
         (c) 2005 The Gale Group
File 160:Gale Group PROMT(R) 1972-1989
         (c) 1999 The Gale Group
File 148:Gale Group Trade & Industry DB 1976-2005/Aug 09
         (c)2005 The Gale Group
File 624:McGraw-Hill Publications 1985-2005/Aug 09
         (c) 2005 McGraw-Hill Co. Inc
File  15:ABI/Inform(R) 1971-2005/Aug 09
         (c) 2005 ProQuest Info&Learning
File 647:CMP  Computer Fulltext 1988-2005/Jul W4
         (c) 2005 CMP Media, LLC
File 674:Computer News Fulltext 1989-2005/Aug W1
         (c) 2005 IDG Communications
File 696:DIALOG Telecom. Newsletters 1995-2005/Aug 08
         (c) 2005 Dialog
File 369:New Scientist 1994-2005/May W5
         (c) 2005 Reed Business Information Ltd.

Set     Items   Description
S1   14383238   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
              OR CATEGORIES OR CLASS OR CLASSES OR GROUP? ? OR CLUSTER? ? -
             OR FAMILY OR FAMILIES OR COLLECTION? ? OR DOMAIN? ?
S2      68095   (FREQUEN???? OR OCCURR? OR INCIDENCE? ? OR HOW()OFTEN OR A-
             PPEAR?)(5N)(TERM? ? OR WORD? ? OR KEYWORD? ? OR ELEMENT? ?)
S3        455   S2(5N)(DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? -
             OR RECORD? ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMAT-
             ION OR CONTENT? ?)(5N)S1
S4    5516383   RATIO? ? OR PERCENT???? OR PROPORTION??
S5         31   S3(100N)S4
S6         20   RD (unique items)
S7         14   S6 NOT PY=2000:2005
S8    4803327   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
              OR CATEGORIES
S9        245   S2(5N)(DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? -
             OR RECORD? ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMAT-
             ION OR CONTENT? ?)(5N)S8
S10       184   RD (unique items)
S11     71960   (CLASSIF? OR CATEGORIZ? OR CATEGORIS?)(5N)(DOCUMENT? ? OR -
             ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? OR RECORD? ? OR PAPER? ? -
             OR MANUSCRIPT? ? OR DATA OR INFORMATION OR CONTENT? ?)
S12        17   S9(50N)S11
S13        17   S9(100N)S11
S14        11   RD (unique items)
```

**14/3,K/1     (Item 1 from file: 275)**
DIALOG(R)File 275:Gale Group Computer DB(TM)

02584919     SUPPLIER NUMBER: 83374389     (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Info-Strainer. (Proto Type).(Brief Article)**
Technology Review (Cambridge, Mass.), 105, 2, 18(1)
March, 2002
DOCUMENT TYPE: Brief Article     ISSN: 1099-274X     LANGUAGE: English
RECORD TYPE: Fulltext
WORD COUNT:   155    LINE COUNT:   00016

StreamLogic's program monitors constantly changing content sources such as discussion groups, newswires and stock quotes and **categorizes** their **information** by **topic** , according to the **frequency** of certain **words** or **word** pairs. It then strains this **categorized** **content** through a mathematical filter; when content matching a preset pattern emerges, the system issues an alert or...

**14/3,K/2     (Item 2 from file: 275)**
DIALOG(R)File 275:Gale Group Computer DB(TM)

02463090     SUPPLIER NUMBER: 68876634     (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Tahoe Makes It Easier To Find And Share Documents -- Upcoming Microsoft portal platform needs tighter application integration.(Product Development)**
Feibus, Andy
InformationWeek, 86
Jan 8, 2001
ISSN: 8750-6874     LANGUAGE: English     RECORD TYPE: Fulltext; Abstract
WORD COUNT:   1201    LINE COUNT:   00099

...    documents on the site, whether published or in progress.
Once published, a portal coordinator is responsible for **categorizing** the **information** . A **document** can be placed in more than one category and a "best bet" category can also be chosen...

...properties as well, including a description and a set of search keywords that will guide users to **documents** , even if the **category** name doesn't **appear** on the list of **words** being sought.
The most interesting part of Tahoe is the index/search engine and its support for...

**14/3,K/3     (Item 3 from file: 275)**
DIALOG(R)File 275:Gale Group Computer DB(TM)

02403857     SUPPLIER NUMBER: 62284974     (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Beyond the Numbers.(Technology Information)**
SULLIVAN, DAN
Intelligent Enterprise, 3, 6, 36
April 10, 2000
LANGUAGE: English     RECORD TYPE: Fulltext; Abstract
WORD COUNT:   3211    LINE COUNT:   00272

...    lets you identify relations such as age, professional status, dependency, shared identity, origin, and family relationships.
The **categorization** tool assigns **documents** to preexisting categories or themes. A training phase is required to create the categories tailored to the...

...the category scheme to identify relevant descriptors and associated vocabulary statistics. You then use these statistics to **classify** the **contents** of the **document** collection. The output of the **categorization** tool is the weighted sum of all the different vocabulary items in the **document** . Weights take into account the relative **frequency** of **terms** in the individual **categories** vs. the entire training set, so that **words** that occur **frequently** in a particular **category** have greater weight than more evenly distributed ones.

The summarization tool extracts sentences that are most relevant...

**14/3,K/4 (Item 4 from file: 275)**
DIALOG(R)File 275:Gale Group Computer DB(TM)
(c) 2005 The Gale Group. All rts. reserv.

02129387 SUPPLIER NUMBER: 20080290 (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Two ways to take stock. (Raosoft's EZSurvey 97 and SPSS's TextSmart Web-based survey software) (Software Review)(Evaluation)**
Simon, Barry
PC Magazine, v17, n1, p73(1)
Jan 6, 1998
DOCUMENT TYPE: Evaluation ISSN: 0888-8507 LANGUAGE: English
RECORD TYPE: Fulltext; Abstract
WORD COUNT: 893 LINE COUNT: 00070

... tool to choose a category. You can also highlight replies and associated keywords. Also featured are a **categories** plot and bar charts of **category** and **word** **frequencies** . You can save the **categorized** **data** in tab-delineated format or in a format that the SPSS statistical-analysis programs can use.

On...

**14/3,K/5 (Item 1 from file: 16)**
DIALOG(R)File 16:Gale Group PROMT(R)
(c) 2005 The Gale Group. All rts. reserv.

06574471 Supplier Number: 55497341 (USE FORMAT 7 FOR FULLTEXT)
**Mining Meets the Web.**
ZORN, Peggy; EMANOL, Mary; MARSHALL, Lucy; PANEK, Mary
Online, v23, n5, p17
Sept-Oct, 1999
Language: English Record Type: Fulltext Abstract
Document Type: Magazine/Journal; Trade
Word Count: 4166

... Data mining models fall into three basic categories: classification, clustering, and associations and sequencing (see Figure 1).

* **Classification** --involves analyzing **data** and assigning it to predefined concept **categories** or "tags," based on predefined rules. Automatically assigning controlled vocabulary **terms** to **records** based on **word** **occurrence** is an example of classification.

* Clustering--similar to classification in that different concept categories are identified through...

**14/3,K/6 (Item 1 from file: 148)**
DIALOG(R)File 148:Gale Group Trade & Industry DB
(c)2005 The Gale Group. All rts. reserv.

0017099772 SUPPLIER NUMBER: 117425351 (USE FORMAT 7 OR 9 FOR FULL TEXT)
**The transformation of document capture.**
Samat, Sameer
Advanced Imaging, 19, 5, 18(3)
May, 2004

ISSN: 1042-0711      LANGUAGE: English      RECORD TYPE: Fulltext
WORD COUNT:   1663    LINE COUNT:  00143

...     the categories represent very similar concepts, determining the right rules manually becomes quite difficult. Lastly, if the **documents** being **classified** are of varying **genres** , the vocabulary and **frequency** of individual **words** will tend to vary from **document** to **document** . Consider the differences between a Lease Agreement and a newspaper article discussing the recent changes in lease...

**14/3,K/7     (Item 2 from file: 148)**
DIALOG(R)File 148:Gale Group Trade & Industry DB
(c)2005 The Gale Group. All rts. reserv.

06504605     SUPPLIER NUMBER: 14174181    (USE FORMAT 7 OR 9 FOR FULL TEXT)
**Indexing - the key to retrieval.**
Wiggins, Bob
Document Image Automation, v13, n2, p13(3)
Summer, 1993
ISSN: 1054-9692      LANGUAGE: ENGLISH      RECORD TYPE: FULLTEXT
WORD COUNT:   1868    LINE COUNT:  00156

...     which it is stored. Catalogue information, as exemplified above, provides the basic data for lists of substantive **documents** and for indexes to them; for example, useful **subject** search **terms** may **appear** in the titles.

**Classification**

While a catalogue **records** the base descriptive information, which may be all that is required to deal with enquiries on the...

...together of items. One has in effect to determine to which conceptual 'pigeon hole' a piece of **information** is assigned. A file **classification** devised as part of a **records** management programme is an example of such a method of organisation.

The file classification and retention schedules...

**14/3,K/8     (Item 1 from file: 15)**
DIALOG(R)File  15:ABI/Inform(R)
(c) 2005 ProQuest Info&Learning. All rts. reserv.

02886555  826457481
**Implementing a Taxonomy Solution**
Lederman, Paula
AIIM E - Doc Magazine  v19n2  PP: 25-26  Mar/Apr 2005
JRNL CODE: EDOC
WORD COUNT: 1152

...TEXT: user friendliness, visualization of the relationship, and easy means by point and click to navigate these hierarchies

**Categorizing  Documents**

The next category of software that is essential to applying taxonomy is the tool used for **categorization** . Ideally, when a new **document** is created or added to a repository or network, the **document** can be passed through software and based on **word** **frequencies** and relationships the **categorizer** can assign the **document** to various **categories** (either subject, language, document types, source, or others). This intelligent categorization is usually presented as a best...

...a user should be able to enter a keyword, and have the categories automatically assigned to the **document** .

Searching Usinq Categories

A **classified** search allows for more comprehensive and precise searching, without dependence of key words and variations in language...

**14/3,K/9 (Item 2 from file: 15)**
DIALOG(R)File 15:ABI/Inform(R)
(c) 2005 ProQuest Info&Learning. All rts. reserv.

02108241 66589040
**Tahoe makes it easier to find and share documents**
Feibus, Andy
Informationweek n819 PP: 86-88 Jan 8, 2001
ISSN: 8750-6874 JRNL CODE: IWK
WORD COUNT: 1147

...TEXT: documents on the site, whether published or in progress.

Once published, a portal coordinator is responsible for **categorizing** the **information** . A **document** can be placed in more than one category and a "best bet" category can also be chosen...

...properties as well, including a description and a set of search keywords that will guide users to **documents** , even if the **category** name doesn't **appear** on the list of **words** being sought.

The most interesting part of Tahoe is the index/search engine and its support for...

**14/3,K/10 (Item 3 from file: 15)**
DIALOG(R)File 15:ABI/Inform(R)
(c) 2005 ProQuest Info&Learning. All rts. reserv.

01331706 99-81102
**An exploration of the espoused organizational cultures of public accounting firms**
Holmes, Scott; Marsden, Stephen
Accounting Horizons v10n3 PP: 26-53 Sep 1996
ISSN: 0888-7993 JRNL CODE: ACH
WORD COUNT: 9300

...TEXT: scores "1" on the participation category, and so on for the other ten content categories.

The eventual **content** dictionary consisted of 59 separate words describing the 11 **content categories** . Based on the **frequency** that key **words appear** within the 11 **content categories** , firms were **classified** into one of four "ideal" culture types-elite, leadership, meritocratic or collegial. Several of the 11 content...

**14/3,K/11 (Item 1 from file: 647)**
DIALOG(R)File 647:CMP Computer Fulltext
(c) 2005 CMP Media, LLC. All rts. reserv.

01229463 CMP ACCESSION NUMBER: IWK20010108S0035
**Tahoe Makes It Easier To Find And Share Documents - Upcoming Microsoft portal platform needs tighter application integration**
Andy Feibus
INFORMATIONWEEK, 2001, n 819, PG86
PUBLICATION DATE: 010108
JOURNAL CODE: IWK LANGUAGE: English
RECORD TYPE: Fulltext

SECTION HEADING: TECH ANALYZER
WORD COUNT: 1116

... documents on the site, whether published or in progress.

Once published, a portal coordinator is responsible for **categorizing** the **information** . A **document** can be placed in more than one category and a "best bet" category can also be chosen...

...properties as well, including a description and a set of search keywords that will guide users to **documents** , even if the **category** name doesn't **appear** on the list of **words** being sought.

```
File 347:JAPIO Nov 1976-2005/Apr(Updated 050801)
         (c) 2005 JPO & JAPIO
File 350:Derwent WPIX 1963-2005/UD,UM &UP=200550
         (c) 2005  Thomson Derwent

Set     Items   Description
S1    1335227   THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR CATEGORY?
              OR CATEGORIES OR CLASS OR CLASSES OR GROUP? ? OR CLUSTER? ? -
             OR FAMILY OR FAMILIES OR COLLECTION? ?
S2    1356994   FREQUEN???? OR OCCURR? OR INCIDENCE? ? OR HOW()OFTEN OR AP-
             PEAR?
S3        405   S2(5N)(TERM? ? OR WORD? ? OR KEYWORD? ? OR ELEMENT? ?)(5N)-
             (DOCUMENT? ? OR ARTICLE? ? OR PAGE? ? OR WEBPAGE? ? OR RECORD?
              ? OR PAPER? ? OR MANUSCRIPT? ? OR DATA OR INFORMATION OR CON-
             TENT? ?)(5N)(S1 OR DOMAIN? ?)
S4    1165935   RATIO? ? OR PERCENT???? OR PROPORTION??
S5         22   S3 AND S4
S6        289   S3 AND IC=G06F
S7         58   S6 AND AC=US/PR
S8         33   S7 AND AY=(1963:1999)/PR
S9        125   S6 AND PY=1963:1999
S10       132   S8:S9
S11        22   S5
S12       123   S10 NOT S11
S13        33   S12 AND (THEME? ? OR TOPIC? ? OR SUBJECT? ? OR GENRE? ? OR
             CATEGORY? OR CATEGORIES)
S14        90   S12 NOT S13
```

**13/5/2 (Item 2 from file: 347)**
DIALOG(R)File 347:JAPIO

06226019 **Image available**
INFORMATION SORTING METHOD, DEVICE AND SYSTEM

PUB. NO.: 11-167581 [JP 11167581 A]
PUBLISHED: June 22, 1999 ( **19990622)**
INVENTOR(s): HARA MASAMI
KITANI TSUYOSHI
APPLICANT(s): NTT DATA CORP
APPL. NO.: 09-334309 [JP 97334309]
FILED: December 04, 1997 (19971204)
INTL CLASS: **G06F-017/30 ; G06F-007/24**

ABSTRACT

PROBLEM TO BE SOLVED: To provide an information sorting device which can sort the texts with high accuracy.

SOLUTION: An information sorting device 1 includes a text input part 11, a word processing part 12, a vector processing part 13, a learning feature vector set file 14, a similarity processing part 15, a **category** decision part 16 and an external or internal **document data** base 17. The part 12 calculates the importance of **category** of every **word** that is extracted from a learning text based on both number of **appearance** and **category frequencies** of the **word** . The part 15 calculates the similarity of **words** based on the learning feature vector, the learning feature vector set and the sorting object text feature vector which are calculated based on the importance of words calculated at the part 12. The part 16 decides a prescribed number of corresponding **categories** as the **categories** of the sorting object texts based on the similarity having the largest calculation value. Then the sorting object texts sorted in each **category** are stored in the" data base 17.

**13/5/3 (Item 3 from file: 347)**
DIALOG(R)File 347:JAPIO

06202335 **Image available**
DEVICE AND METHOD FOR KEYWORD WEIGHT GENERATION AND PROGRAM STORAGE MEDIUM

PUB. NO.: 11-143892 [JP 11143892 A]
PUBLISHED: May 28, 1999 ( **19990528)**
INVENTOR(s): OKAMOTO AOSHI
YUGAMI NOBUHIRO
MATSUMOTO SHUNJI
APPLICANT(s): FUJITSU LTD
APPL. NO.: 09-305167 [JP 97305167]
FILED: November 07, 1997 (19971107)
INTL CLASS: **G06F-017/30**

ABSTRACT

PROBLEM TO BE SOLVED: To make it possible to generate an appropriate weight of a keyword in a keyword weight generation device for generating weight of a keyword appearing in a document.

SOLUTION: This device comprises a first calculation means 10 that obtains statistical information on each keyword appearing in a **document** by referring to a **document** database and calculates weight of each **keyword**

St. Leger, Geoffrey

#10088895_2

Access DB# 161637

# SEARCH REQUEST FORM

## Scientific and Technical Information Center

(39)

Requester's Full Name: Gwen Liang Examiner # : 79180 Date: 8-5-05
Art Unit: 2162 Phone Number 30~~8~~ x24038 Serial Number: 10/888,895
Mail Box and Bldg/Room Location: RND 3B11 Results Format Preferred (circle): [✓ PAPER] DISK E-MAIL

**If more than one search is submitted, please prioritize searches in order of need.**

**************************************************************************************

Please provide a detailed statement of the search topic, and describe as specifically as possible the subject matter to be searched. Include the elected species or structures, keywords, synonyms, acronyms, and registry numbers, and combine with the concept or utility of the invention. Define any terms that may have a special meaning. Give examples or relevant citations, authors, etc, if known. Please attach a copy of the cover sheet, pertinent claims, and abstract.

Title of Invention: Method of Thematic Classification of Documents ...

Inventors (please provide full names): BIETTRON, Laurent ; PALLU, Frederic ; TRICOT, Sylvie

Earliest Priority Filing Date: 9-24-999 *Assignee = France Telecom

***For Sequence Searches Only* Please include all pertinent information (parent, child, divisional, or issued patent numbers) along with the appropriate serial number.**

Background & Concept: (See "CON" pages)

7 Claim = 14 (focus on 14-5-2, 14-10, 14-11, 14-12)
(See "CLM" pages)

For arguments for claim 14 (See "Remarks" pages

Prior art = US 5625767 (Bartell et al.)
(Wilbur) "An analysis of statistical term strength and its use ..."
(as 2 pages attached, marked "REF")

**************************************************************************************

| | | |
|---|---|---|
| **STAFF USE ONLY** | **Type of Search** | **Vendors and cost where applicable** |
| Searcher: Geoffrey St. Leger | NA Sequence (#)____ | STN ____ |
| Searcher Phone #: 23540 | AA Sequence (#)____ | Dialog ✓ |
| Searcher Location: 4B31 | Structure (#) ____ | Questel/Orbit ____ |
| Date Searcher Picked Up: 8/8/5 | Bibliographic ____ | Dr.Link ____ |
| Date Completed: 8/9/5 | Litigation ____ | Lexis/Nexis ____ |
| Searcher Prep & Review Time: 60 | Fulltext ✓ | Sequence Systems ____ |
| Clerical Prep Time: ____ | Patent Family ____ | WWW/Internet ____ |
| Online Time: 200 | Other ____ | Other (specify)____ |

A METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE

The present invention relates to a method of thematically classifying documents and intended in particular for setting up or updating thematic databases, in particular for a search engine.

The invention also relates to a module for thematically classifying documents, and to a search engine fitted with such a thematic classification module.

At present, two main computer tools are known for searching documents on a computer network such as the Internet, for example.

Background

These tools are search engines and guides.

A search engine is a tool that serves to extract the words or terms that are most representative of information, mainly in the form of text, and to store them in a database, also known as an "index" base.

Such index bases are generally updated relatively frequently.

In response to a request made by a user, the same tool scans through the index bases in order to identify the terms which are most relevant relative to those of the request, and then to sort the information obtained in return.

The other technique for searching for documents on a computer network consists in using a guide. That tool proposes searches by category, with document pages being classified manually by researchers.

Those types of tool present various drawbacks.

Firstly, search engines do not propose classifying document pages by category. The pages provided in response to a request are not typified. Thus, ambiguous requests can give rise to a very wide variety of responses that are perceived by the user as noise.

CON (1/3)

In contrast, guides provide a user with responses that are typified, i.e. that relate to the same theme(s) as the request.

However manually classifying document pages involves high creation and updating costs while allowing only a limited number of pages to be indexed. Consequently, some requests do not obtain any response.

Background

The object of the invention is to mitigate the drawbacks of search engines and of guides.

The invention thus provides a method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, the method being characterized in that it comprises the following steps:

- selecting a sample of documents representative of each theme;
- identifying within the selected documents elements that are characteristic of each theme;
- allocating a coefficient to each identified element, which coefficient is representative of the relevance of said element relative to the corresponding theme;
- for each document to be classified, identifying said theme-characterizing elements that are contained in the document for each of the themes, and for each theme corresponding to the documents, using the coefficients allocated to said elements to calculate the value of a characteristic representative of the relevance of that theme for the document, in order to decide whether or not the document relates to the theme, said identification and calculation steps being performed automatically for each document downloaded from a computer network;
- classifying the downloaded documents as a function of the themes with which they deal; and
- storing the documents classified thematically in databases that can be interrogated on the basis of themes contained in a request;

CON (2/3)

Concept

# ABSTRACT

## A METHOD OF THEMATICALLY CLASSIFYING DOCUMENTS, A THEMATIC CLASSIFICATION MODULE, AND A SEARCH ENGINE INCORPORATING SUCH A MODULE

This method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, comprises the steps of selecting documents representative of each theme, identifying within the selected documents, elements that are characteristic of each theme, allocating a coefficient to each identified element, the coefficient being representative of the relevance of the element relative to the corresponding theme, and for each document for classification, identifying the elements characteristic of each theme contained in the document and, for each theme corresponding thereto, using the coefficients allocated to the elements to calculate the value of a characteristic representative of the relevance of the theme for the document, in order to decide whether or not the document relates to the theme.

CON(3/3)

AMENDMENTS TO THE CLAIMS:

This listing of claims will replace all prior versions, and listings, of claims in the application:

LISTING OF CLAIMS:

1-13. (canceled)

14. (currently amended) A method of thematically classifying documents, in particular for making up or updating thematic databases for a search engine, the method comprising the following steps:

14-1 - manually and/or automatically selecting a sample of documents representative of each theme;

14-2 - automatically identifying within the selected documents elements that are characteristic of each theme;

14-3 - automatically allocating a coefficient to each identified element, which coefficient is representative of the relevance of said element relative to the corresponding theme;

14-4 - downloading documents from a computer network;

14-5 - 14-5-1 for each downloaded document to be classified, identifying said theme-characterizing elements that are contained in the document for each of the themes, 14-5-2 and for each theme corresponding to the elements, using the coefficients allocated to said elements to calculate [[the]] a characteristic value ~~of a characteristic~~ representative of the relevance of that theme for the document, in order to decide whether or not the document

CLM 1/2

14-5-3

relates to the theme, said theme-characterizing elements identification and calculation steps being performed automatically for each document downloaded from [[a]] the computer network;

14-6 - automatically classifying the downloaded documents as a function of the themes with which they deal; [[and]]

14-7 - automatically storing the documents classified thematically in databases that can be interrogated on the basis of themes contained in a request; and

14-8 - making the databases available to users who interrogate the databases on the basis of themes contained in a request;

14-9 and the step of allocating said coefficient to each identified element comprises the following steps for each theme:

* 14-10 - automatically calculating [[the]] a frequency of the element in the selected documents relating to the theme;

* 14-11 - automatically calculating [[the]] a frequency of the element in the selected documents that do not relate to the theme; and

* 14-12 - automatically calculating the ratio of the calculated frequencies.

15. (previously presented) A method according to claim 14, further comprising the step of automatically sorting themes in a theme tree structure in decreasing order of coefficients.

CLM (2/2)

Claims 14-26 were rejected as unpatentable over WILBUR et al. ("An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts"). Reconsideration and withdrawal of the rejection are respectfully requested.

argument
cl. 14

WILBER et al. defines the "strength" of a word and uses this strength to remove weak words from a set of words training a categorization learning system. However, the "strength" defined in WILBER et al. does not correspond to the claimed characteristic value representative of the relevance of a theme for a document. Two documents that are relevant to each other in the reference do not necessarily belong to the same theme. Since "strength" in the reference refers to relevance between documents, the strength does not necessarily relate to the theme of a document. Further, as is apparent from the example at pages 212-213 the strength is a mean value not related to a theme. If one considers that d1, d2, x1, and x2 belong to a first theme, that d1, d2, and x2 belong to a second theme, and that d3 and x3 belong to a third theme, then the calculated strength is not related to these themes since is a mean value. Accordingly, the reference does not disclose the step of or means for calculating a characteristic value representative of the relevance of that theme for the document, and thus these claims avoid the rejection under §103.

Remarks (1/2)

In addition, WILBUR et al. does not disclose the step of or means for automatically calculating a frequency of the element in the selected documents relating to the theme, automatically calculating a frequency of the element in the selected documents that do not relate to the theme, and automatically calculating the ratio of the calculated frequencies. The Official Action points to Figure 2 and formula 18 on page 219. However, the ratio disclosed therein is not the same ratio as claimed. The reference discloses a ratio of words removed to total words, not a ratio of the frequency of the element in the documents relating to the theme to the frequency of the element in the documents not related to the theme. There is no suggestion in the reference to find this ratio and the claims thereby further avoid the rejection under §103.

Claims 23-25 have been amended to recite a method, instead of a use.

In view of the present amendment and the foregoing remarks, it is believed that the present application has been placed in condition for allowance. Reconsideration and allowance are respectfully requested.

The Commissioner is hereby authorized in this, concurrent, and future replies, to charge payment or credit any

Remarks (2/2)

# ADONIS - Electronic Journal Services

Requested by

Adonis

| | |
|---|---|
| Article title | An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts |
| Article identifier | 0010482596000252 |
| Authors | Wilbur_W_J Yang_Y |
| Journal title | Computers in Biology and Medicine |
| ISSN | 0010-4825 |
| Publisher | Pergamon |
| Year of publication | 1996 |
| Volume | 26 |
| Issue | 3 |
| Supplement | 0 |
| Page range | 209-222 |
| Number of pages | 14 |
| User name | Adonis |
| Cost centre | |
| PCC | $20.00 |
| Date and time | Tuesday, November 09, 2004 10:50:26 AM |

REF(1/2)

Pergamon

Comput. Biol. Med. Vol. 26, No. 3, pp. 209–222, 1996
Copyright © 1996 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0010–4825/96 $15.00 + 0.00

0010-4825(95)00055-0

# AN ANALYSIS OF STATISTICAL TERM STRENGTH AND ITS USE IN THE INDEXING AND RETRIEVAL OF MOLECULAR BIOLOGY TEXTS

W. John Wilbur*† and Yiming Yang‡

†National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, U.S.A.; and ‡Section of Medical Information Resources, Mayo Clinic/Foundation, Rochester, MN 55905, U.S.A.

(*Received* 20 *October* 1995; *and in revised form* 26 *October* 1995)

**Abstract**—The biological literature presents a difficult challenge to information processing in its complexity, diversity, and in its sheer volume. Much of the diversity resides in its technical terminology, which has also become voluminous. In an effort to deal more effectively with this large vocabulary and improve information processing, a method of focus has been developed which allows one to classify terms based on a measure of their importance in describing the content of the documents in which they occur. The measurement is called the strength of a term and is a measure of how strongly the term's occurrences correlate with the subjects of documents in the database. If term occurrences are random then there will be no correlation and the strength will be zero, but if for any subject, the term is either always present or never present its strength will be one. We give here a new, information theoretical interpretation of term strength, review some of its uses in focusing the processing of documents for information retrieval and describe new results obtained in document categorization. Copyright © 1996 Elsevier Science Ltd.

Molecular biology  Stop terms  Text retrieval  Text classification
Linear least squares fit  Weight  Strength  Vector model  Bayesian model

## 1. INTRODUCTION

As the complexity and volume of biological literature has continued to grow at an accelerating pace, methods of processing and making available to researchers the literature relevant to their research interests has become more difficult. Numerous proposals of AI techniques and knowledge based systems (see, e.g. [1]) have been made to deal with this problem, but unsolved problems in natural language processing have rendered such approaches of limited practicality up to the present time. In our judgment the automatic methods based on keyterms and a vector or probabilistic model of retrieval remain the most effective methods of text processing and analysis for very large databases.

The MEDLINE® database [2, 3] maintained by the National LIbrary of Medicine contains over seven million records. In the G5 (genetics) subset alone there are over 1.2 million records. The G5 subset represents an area of particular interest for us and we have processed it with a resultant vocabulary list of over 1.2 million technical terms or strings from free text and over 200,000 MESH®/qualifier combinations. MESH (Medical Subject Headings [4]) is a controlled vocabulary of key terms and phrases in the areas of biology and medicine that are used in indexing the MEDLINE database. The MESH terms are organized as a number of trees with increasing specificity of terms as one moves down the branches towards the leaves. Qualifiers are subheadings denoting broad categories such as 'etiology', 'treatment' or 'virology' that are used to increase the specificity of a MESH term. While the MESH/qualifier terms are constructed in such a manner as to be generally useful, they provide only very limited coverage in the highly specialized subareas of molecular biology. When compared with the MESH, however,

[*] Author to whom correspondence should be addressed.

KZF( 2/2)

0467393 94-20676
**Individual's interactive newspaper hits stride**
Porter, Patrick L
MASS HIGH TECH (Watertown, MA, US), V12 N3 s1 p2
PUBL DATE: 940124
JOURNAL CODE: MAHT DOCUMENT TYPE: Newspaper article
WORD COUNT: 5,015
DATELINE: Cambridge, MA, US

TEXT:

MHT: What led you to the idea of a personalized interactive newspaper?

AMRAM: I had started to look at this notion of the knowledge worker and how one might help them filter and synthesize information. I saw that traditional on-line databases that were supposed to provide access to information were hard to use and didn't really address the end-user needs--even I had been a frustrated user of Dialog and Dow Jones News Retrieval. At Aegis we had looked at making some investments in companies like Delphi (an on-line service) but we decided not to because we saw that their market was limited to the PC professional and enthusiast.

MHT: You have to enjoy rummaging around in a non-user-friendly environment. I think people using those systems have to get a certain kick out of the problem-solving needed to find the information they're looking for. That has to be part of their reward--figuring out the system--as opposed to a business person who is really only interested in finding an answer.

AMRAM: Exactly. Those on-line systems were just too hard to use. I remember a story Dan Bricklin tells about his invention of the PC-based electronic spreadsheet where he went to a finance professor at Harvard and told him about his idea and the professor said, "What's the big deal? Anything you could do with a PC spreadsheet you could do with a financial modeling language on a mainframe--like Express."

But the point was that when you could package the power of something that takes weeks of training on a mainframe into something that an end user can learn in 15 or 20 minutes, you can create a whole new medium. So the question is how can you take the power of these electronic databases, this ever growing content that is available on on-line databases, that is more timely, that provides you this capability to selectively filter it because once it's electronic you can search it and filter it (as opposed to a static medium) and package it for the end user? At the same time some new technologies were coming in like fax and desktop publishing and electronic mail networks. And those things created the infrastructure for packaging the information and disseminating it very rapidly.

So I went around to some professors I knew at MIT and started looking at technologies that could be applied to the filtering. I found out about some of the work that is being done at the Media Lab and their concept of the personal newspaper. Their model was a little different in that you were broadcasting the information to the PC and the filtering was being done locally. They had this concept, though, that the newspaper should learn about you as you used it, so the longer you looked at the article it deduced that it was more relevant to you. But it wasn't particularly practical in that people might got to the bathroom and leave it on their screen and it would think the article was relevant. They I didn't really build the algorithms underneath it. They had more of a user-interface prototype.

MHT: There is some work at the Media Lab by Prof. Patti Maes. She has gone a couple of steps beyond that by developing learning interface agents.

But first I would like to hear about Dr. Gerard Salton and how you ran into SMART and then maybe we could talk about how SMART is similar to or different from what Patti Maes is doing.

AMRAM: Okay. Through my investigations at MIT I also talked to David Gifford there who was in the computer science lab. And he built this computer information system in Boston that was broadcasting information to PCs and doing the filtering, He also built an information service that was based on a profile of the user and was distributing information on email. But it was a basic key-word-based system search. So he was more focused on the architecture but not so much the technology for making the searches more effective. Anyway, I started asking him about papers and who was doing the state-of-the-art work in text understanding and text retrieval and all the paths seemed to lead to Salton at Cornell. He was the person who was doing the key work in that area.

So I called Salton at Cornell and said, "I've got this business concept and I think your technology could be applied I went up to Cornell and met with him and he said it sounded interesting but there had been a lot of entrepreneurs who had tried to commercialize his technology but nothing had come of it. He spent a lot of time with one of them and they couldn't get funding and it ended up being a lot of wasted time.

MHT: What was wrong with their model?
AMRAM: They were just trying to take his technology and sell it as text retrieval software. They couldn't get the company off the ground. But I convinced him over time through my persistence that this was going to be real and he agreed to provide the license. At the same time I was trying to pull together a team so he turned me on to one of his former Ph.D. students who was involved in building SMART at Cornell. The guy's name was Harry Wu.

And at the same time there was Jacques Bouvard from Honeywell who had worked with Harry building their knowledge-based system. So the two of them joined from the technical side. And then I was looking for someone who had a publishing background to join the team and that was John Zahner. So that was the initial team. We incorporated in early '89. By then Aegis decided they weren't interested in funding it.

The way these partnerships work is you've got to build consensus. There were four partners, two of them decided this was not worth investing in and they couldn't go forward. So I had to raise funding elsewhere.

MHT: Where did you get the money?

AMRAM: We raised $100,000 of seed capital, half from my family and friends...
MHT: And they're still your friends?

AMRAM: (laughing) Yes. A lot of people advised me not to do that because of the risk of destroying those relationships, but luckily it worked out okay. Also, I ran across Ed Fredkin...

MHT: That name sounds very familiar.

AMRAM: He's fairly well known in the Boston high-tech community now but he was a college dropout of Cal Tech and started a company called

Information International in the '60s that was building computer pre-press systems and color separations for newspapers. He took that public and sold the company and then went to MIT and actually built their computer science department and lab. Even though he was a college dropout he was the director of lab at computer science. He and Minsky built the AI Lab at MIT. He was interested in the concept because it could mine computer-based intelligence and publishing--two angles of what he's done. So he invested $50,000 and gave us an office and helped us get off the ground.

At that point we had the technology, we had a core team, and now we had to go out and get some content providers to license us the information, because one of the questions that early on was a big risk was how would the people who own the content view this? As a new distribution channel? Would it cannibalize their core revenue base? Would they license us? Would customers be willing to pay a premium for filtered, personalized information that is uniquely tailored to them?

MHT: Did you have market research or just a gut feeling that it would work because potential users had hit information overload?

AMRAM: Well it was both a gut feeling and some informal surveys of many people. I had been in high tech myself and I knew that if someone could give me the news that was relevant to me--and I was overwhelmed with stacks of magazines I couldn't go through and that weren't timely--then it would be worth the money. I did some informal market research in Boston where I surveyed executives randomly, but it wasn't hundreds of people in formal focus groups. I didn't have the resources to do that. Also, all the fundamental trends were in favor. If it wasn't the right idea in '89--as more and more content was being published the amount of information was doubling every three years, more and more raw material was there, product cycles were shrinking, markets were getting global, that meant there was going to be more and more value placed on getting relevant information into the hands of people so they could make more effective decisions. Those fundamental, immovable forces were working in our favor.

MHT: How did you sign your first information providers? Who were they?

AMRAM: The first group of providers were the Business Wire and the PR Newswire which are press release newswires, and UPR and Kyoto. They were the first four that we got at the end of '89.

MHT: Those were relatively easy since they were subscription services, right? I mean, they didn't have a proprietary information base. The PR Newswire sells to all comers. If you want to sign up for the Business Wire you just do it.

AMRAM: They were easier to get. But they also represented a good value because a lot of high-tech companies wanted to see the press releases of their competitors and if you're just waiting for the stuff to appear in the trade press you can wait a long time.

MHT: It doesn't always appear. If an editor decides it isn't worth running, it just disappears.

AMRAM: Exactly. And some details about how your competitors quote their products, pricing and configuration and information may never make it into the trade press. So based on those four services we were able to sign up some early customers, companies like Lotus and Digital, and we sold them a few profiles. That helped us validate that there was a market for it and that at least we could get some content and work the technology, taking this thing out of the university, integrating it and building it up into a workable system. And based on that we were able to secure some additional venture capital.

MHT: From whom?

AMRAM: The first two professional venture firms were Grace Ventures and Venture Capital Fund of New England. We also brought in some additional private investors who were successful entrepreneurs in the Boston area, Mort Goulder who was a founder of Sanders, and Andy Devereaux, one of the founders of American Cable Systems which later was sold to Continental Cablevision, and Ted Johnson, one of the early employees at Digital Equipment. He built their sales and marketing department. So it was a combination of professional venture investors as well as some successful entrepreneurs. We raised about $1 million. That was at the end of '89. Then we rolled out the service in 1990.

MHT: And at some point you started signing a lot more providers and then I think your service became really valuable. I'm not going to read off their names, but how did you get them? It's quite a list.

AMRAM: It's a continuous process and this relates to our strategy and how it works. Once we get a few subscribers and they say, well, this is nice but we're missing these and these services and if you want to give us more valuable information you need a broader pool of sources to draw on. But once we had, for example, Jim Manzi at Lotus reading First, or Esther Dyson or Stewart Alsop, people that were fairly well known, we could go to the content owners and say, "Do you want to get your content to those people?" So as we got more subscribers we had more economies of scale and better distribution channels to go to the content owners. Also, we were able to get some experience once we got a few newsletter publishers to demonstrate to them that this was not cannibalizing their core print subscription base but it was producing incremental revenue for them. And as people were seeing frequent hits in a particular magazine or newsletter they would in some cases actually want to subscribe directly to them. So we were acting as another channel to promote their editorial. We were getting more and more receptivity. And then because we had more and more subscribers we were able to afford to pay more in royalties to the content providers. And because the way we pay royalties is based on usage, the more relevant their information is to our customers the more royalties they get. And the more customers we have obviously the more royalty it means to them because every time an article hits someone's profile from Health News Daily or Cancer Week, that content provider is receiving a royalty for it. Obviously, the more subscribers we have the more royalty that they're going to get. So that kind of feeds on itself. As we get more content providers the service becomes more compelling for the subscribers. The more subscribers we get the more royalty and audience that the content providers are interested in going to.

MHT: How many information providers do you have? I know you search through some astronomical number of articles every day.

AMRAM: We had between 200 and 300 content providers the last time I counted. We feed through over 10,000 articles a day.

MHT: Is there a gigantic mainframe chewing through this?

AMRAM: No. We have some PCs that capture the news on the front end, about 7 or 8 of them. It comes to us in broadcast mode from the real-time wire. We have an FM receiver on the roof or we have leased lines into the news providers. And the PCs basically monitor the flow of news and look for a beginning of story and end of story type of character. They store each story as a file on the local disk on the PC. So each PC can monitor about two live wires. Then there are some sources like the newspapers and the trade periodicals that download once a day or once a week, and there we dial out to some remote computer like the Financial Times host in the U.K. and download tomorrow's issue of the Financial Times electronically as a file. Those stories then get stored on the local disk and they also get sent over a local area network to what we call the system controller which acts as a traffic cop on our network. It does load balancing between multiple computers here. Then the stories go to a story editor. The story editor converts all the formats of the different databases and newswires into a common format that SMART can understand. It looks for where the headline is, where the author is, where the beginning of the text is. If there are tables in the article it marks those as tables so when we do the layout it comes out nicely and so on. The other thin the story editor will do is if the story editor comes in multiple installments throughout the day, frequently there is a flash headline at 10 o'clock and at 10:05 a paragraph gets added. At 10:20 there is more interpretation and analyst comments. The story editor will pull all of those together and make it into a single story--and then passes it to SMART. SMART takes the story and analyzes the content and creates what we call a vector which is a set of concepts and the associated weights of how important those concepts are within that article and takes that representation of the article and matches it against the profile of each user and compares those two vectors comes up with a numerical score, which is how relevant that story is to that profile. It's a number between zero and one. So it's a relative thing as opposed to a yes/no. In a traditional search system you've got a key-word search so if you put in "IBM" you might get a Mike Tyson rape trial article just because one of the jurors works at IBM.

In a key-word-search system there is no way to distinguish between levels of relevance. If you've got two stories, one about IBM and their announcement of the Power/PC chip versus an article that mentions IBM because one of the jurors in a Mike Tyson rape trial article talks about IBM, they both hit the keyword IBM.

MHT: There's no intelligence whatsoever in such a system.

AMRAM: Right. But here we're actually measuring the relevance and normalizing it. So throughout the day as 10,000 articles come in they all are getting ranked and compared against each profile and they are all piling up. For every profile we have an output bin that says "how relevant is that story to that profile?" At about 8:00 p.m. each night we start going through every output bin and for each profile taking the highest ranked 50 or so articles and do a second pass on those. And frequently, like with the Bell Atlantic/TCI merger, that story might come to us from 60 different sources. The Boston Globe covers it, Financial Times covers it, Reuters and Knight Ridder have reports and so on. There are tons of versions.

MHT: What do you do?

AMRAM: That's where we have some content analysis that actually takes all the articles that match the profile for relevance and says, "okay, but are these all talking about the same event?" So some very sophisticated heuristics compare the articles and if they are talking about the same event they decide which version is most appropriate based on the length of the article, the priority of that article and the source. For certain topics certain sources are covering it. We also actually have to look at the last five days of what we sent out to that subscriber because frequently sources get out of phase. Reuters might have covered it yesterday but the Chicago Tribune didn't cover it until a day or two later. Now if the Chicago Tribune version adds significant value and has a different angle because more information is out there and it is going to be significantly different, then we would want to send it. But if not, then we have to eliminate it.

All of that gets done in the second pass which starts at about 8:00 p.m.

MHT: This is all done by SMART without human intervention?

AMRAM: Yes. This is all automated. Now SMART is actually running on a Unix workstation. The front-end data acquisition is happening on PCs. Then the final selections are taken off the workstations by PCs again that run the desktop publishing system and lay it out, fax it out or email it out. We have two services, one is First which goes typically to a corporation or a department in a corporation and it is frequently fed via electronic mail into a local area network so an enterprise or a business unit can share that information across the whole company or business unit so every key manager is focused on their external market and they are institutionalizing current awareness in the business unit.

MHT: It's very helpful in a business setting to share information across a workgroup because then everybody is on the same page. You can work in unison.

AMRAM: That's exactly right. And with Heads Up we have a personalized product for one individual and there we are trying to deal with information overload. We condense it into a page and provide about 15 to 20 briefs, which is very easy for a user to go through in five minutes or less. And then if they want to see more detail they can interactively request that on demand and he fulfill it within 15 to 30 minutes.

MHT: I've been getting Heads Up for the last six months. While it's nice to get the one page overview and be able to request a fax of the entire article, I could worry a bit whether I was getting all the information I need.

AMRAM: That's why we have this notion of relevance feedback where the system learns over time. And in our First service we explicitly ask our subscribers to rate the articles as relevant or not.

MHT: How does that work? because this is the thing that attracted me to Patti Maes' at MIT--her learning interface agents as opposed to the knowledge engineering approach to building a system. I thought yours was more of a knowledge-engineered approach until I started reading documentation that intimated it was constructed by a knowledge engineer, first to provide the questionnaire that users fill out, but that apparently is not the end of it, there is an interaction and the system gets tuned to the individual user.

AMRAM: Exactly. In the First service we have the relevance feedback process where weekly in the first month and monthly thereafter we ask the subscriber to rate the articles as relevant or not. The SMART system is capable of tuning itself based on the articles that you rate relevant. I takes the concepts that appear frequently in the relevant articles and increases their weight and reduces the weight of the concepts that appear frequently in the non-relevant articles. That feedback helps us deal with what we call type 1 errors. It helps you hone in at the relevance but it doesn't help you as much in dealing in what you might have missed because you are looking at what was sent to them. To deal with the other issue we have this thing we call "interactive fulfillments." In First, in addition to sending the 10 articles in full text, we send briefs of the articles that we think almost made it, another 15 articles maybe, and then if they request the full text of any of those articles, we monitor and learn from it. And that's the same principle we use in Heads Up. Depending on which topics you request more information on the system can learn to increase the priority of those topics. So from that standpoint it's a similar concept to what Patti Maes has been doing recently in the area of genetic algorithms and learning agents.

MHT: That's good stuff. To me that's where the technology starts to get very interesting.

We've talked about sources. Now about customers? Who they are? Why do they use it?

AMRAM: We have customers across a range of industries. We started out focusing on the information technology industry as a vertical but we've rolled out into telecommunications, health care, energy, defense, automotive and financial services. So there are a number of industries. They tend to be information intensive industries where there is a rapid rate of change. They tend to be both in sales and marketing, product development, strategic planning and in some cases purchasing. And they tend to use it for different applications. The competitive analysis type people and the marketing people tend to use it to track an industry or their competition. The sales people tend to use it to track their customers. If a customer opens a new plant or announces an expansion or layoffs or whatever, the sales person is attuned to the opportunities and problems in that client base. The purchasing departments use it to track vendors so if a vendor announces bankruptcy or gets in financial trouble, they know to watch out. Similarly in banking we've got loan officers and risk management officers using it to track their portfolio. So from a credit-watch point of view, if you're working for Fleet Bank you might track your major customers and their industries.

We also deliver Heads Up via wireless. In that environment it's a mobile executive who is using it to stay in touch while they're on the road.

MHT: That's available through Motorola? AMRAM: Yes. Embark is their wireless email. It gets delivered into a handheld wireless computer so while you are on the road you can continue to receive news that is tailored to your interests. So the range of people is from the product manager professional up to the CEO and senior executive.

MHT: There is so much talk about the coming information superhighway but the focus is on the physical plant. Yet here you are with a company that is really using the infrastructure available now and you are providing content with business value. I would like to hear your thoughts about the information superhighway. I talk with people about how we should have fiber to the home, or digital transmission over twisted pair or coaxial cable, but once you get past having 500 television and movie channels and 40 home shopping networks, nobody is really too clear about what is going to be pumped across the system that has any value other than as entertainment, news or education.

AMRAM: And that's why people are saying while it's great to have 500 channels, once you've got that how do you navigate through it? People who use the Internet spend a lot of time navigating and surfing it. How do you know what's out there? You've got this massive ocean of information which provides an awful lot of choice, but how do you productively get what you want out of it? And that's where companies are trying to build some interface software to this thing. You're going to need a whole computer, basically, on your set-top to help you navigate through those 500 channels and you'll need some agents, if you will, that know what you interest is, so if you like Robert Redford, for example, you can find his movie without having to scan through 500 channels. In that environment they are looking at it from a consumer entertainment standpoint. But if you look at where most technologies get adopted, I think that first they get adopted in the business environment where there is a real economic benefit from the usage--it makes people more productive and there's a business justification for it. And then as you ride the cost curve and the technology matures it goes to the home market. In the early '80s PCs were strictly a business tool. Even when IBM launched the home PC, the PC Jr. in '84, it was a big flop. But now 10 years later the price point, the education of the consumer and the maturity of the technology has gotten to the point where the PC is very hot in the home market.

We're obviously focusing on the high-end business application because we think that's where it's at for the next several years until the cost comes down. But having the national information highway and infrastructure will be great for us because it will reduce the transmission cost. If you've got a high bandwidth network we can deliver more information that is a lot more friendly and appealing without paying the astronomical cost of sending things over email networks which are a lot more expensive, which forces the price point higher, which limits your market to the high-value-added application.

MHT: What about applying this technology to voice and video data? For example, there's so much broadcast news today - CNN, CNBC, NPR, and scores of others. Is there a way in the future to incorporate it? People are talking about multimedia databases.

AMRAM: The technology to understand the image and to have a profile to look at the image doesn't exist yet. It's really in the research stage. But what you can do is tag the multimedia image with text that describes what the image is about. And you can use a SMART like system which analyzes and understands the text to be the profiling agent, if you will, and so the selection is done based on the text, but then you get presented with the multimedia and the voice. But in terms of understanding the voice or the image we don't have technology that says how do you take the video image of a movie and represent it as to whether that's a movie you like or not? What you could say is this movie has these actors and it's a horror movie or an action adventure or a comedy, and who the director is, etc., and then once you do that you can have an agent that basically looks at those tags, a description of the movie, and then selects it based on your profile. But there's no way to look at the video and represent that against your profile. But you can certainly apply this technology against those textual tags and filter that.

MHT: Tell me about the Dialog relationship.

AMRAM: Basically there are three sides to the relationship. They invested in the company and are providing us some capital. The second side is we are going to take some of the databases that they have already gathered and use them. While we are continually increasing our source pool, they have a large number of databases and by working with them we can have access to a lot of sources that we will deliver to our customers through First and Heads Up. And thirdly we will be developing some new products that they can market to their audience based on our existing technologies and some extensions that we're working on with them. So those are the three aspects of the relationship. But fundamentally they recognize that they have this massive database of information but the technology they have and the model they have is not going to let them get into the broad audience of professional knowledge workers. We are working to marry our technology with the deep pool of sources they have to reach that audience. In addition, they have a large sales, marketing and distribution force that we want to leverage. So it's a nice synergy. They approached us this spring.

COMPANY NAMES: Individual Inc, Cambridge, MA, US, SIC:7375,
CLASSIFICATION CODES: 8302 (Software and computer services); 2130 (Executives)
DESCRIPTORS: Software industry; Electronic publishing; Executives; Interviews; New England
NAMED PERSONS: Amram, Yosi
?